Humboldt Universität zu Berlin

Visualisierung und Analyse multivariater Daten in der gartenbaulichen Beratung - Methodik, Einsatz und Vergleich datenanalytischer Verfahren

DISSERTATION

zur Erlangung des akademischen Grades doctor rerum horticultorum (Dr. rer. hort.)

Landwirtschaftlich-Gärtnerische Fakultät

Stefan Krusche

Prof. Dr. Dr. h. c. Ernst Lindemann

Gutachter: 1. Prof. Dr. E. Thomas

2. Prof. Dr. W. Bokelmann

3. Prof. Dr. E. W. Schenk

eingereicht: 1.3.1999

Datum der Promotion: 16.12.1999

Schlagworte

Statistik, multivariate Verfahren, graphische Verfahren, explorative Statistik, Visualisierung, Dimensionserniedrigung, Biplots, graphische Modelle, Klassifikations- und Regressionsbäume, formale Begriffsanalyse, Kennzahlen, betriebsbegleitende Untersuchungen, Gartenbau, Beratung

Keywords

Statistics, Multivariate methods, Graphical analysis, Exploratory statistics, Visualisation, Reduction of dimensionality, Biplots, Graphical models, CART, Formal concept analysis, Microeconomic indicators, On-site investigation of crop production, Horticulture, Consulting, Extension

Abstract

In order to interpret large data sets in the context of consultancy and extension in horticulture, this thesis attempts to find ways to visually explore horticultural multivariate data, in order to obtain a concise description and summary of the information available in the data and moreover develop possibilities to interactively analyse survey data.

The thesis is part of an exploratory data analysis which analyses data without making specific model assumptions, is predominantly descriptive, analyses data step by step in a highly interactive setting, and makes full use of all kinds of graphical displays.

The methods used comprise various dimensionality reduction techniques (principal components analysis, correspondence analysis, multidimensional scaling), biplots, the multivariate analysis of grouped data (procrustes rotation and groupwise principal components), graphical models, CART, and line diagrams of formal concept analysis. In addition, further graphical methods are used, like e.g. trellis displays.

Data from an on-site investigation of the production process of Cyclamen in 20 nurseries and from the microeconomics indicators of 297 growers in Germany (so called Kennzahlen) from the years 1992 to 1994 are used to demonstrate the analytical capabilities of the methods used. The data present a perfect example of unperfect data, and therefore represent the majority of the data sets that horticultural consultancy has to work with. Thus, it becomes clear, that despite the variety of results, which helps to enhance the understanding of the data at hand, not only the complexity of the processes observed, but also the low data quality make it fairly difficult to arrive at clear cut conclusions.

The most helpful tools in the graphical data analysis are biplots, hierarchical line diagrams and trellis displays. Finding an empirical grouping of objects is best solved by classification and regression trees, which provide both, the data segmentation, and an intuitively appealing visualisation and explanation of the derived groups. In order to understand multivariate relationships better, discrete graphical models are well suited.

The procedures to carry out a number of the methods which cannot be found in general statistics packages are provided in the form of Genstat codes.

Zusammenfassung

Ausgangspunkt der vorliegenden Arbeit ist die Suche der gartenbaulichen Beratung nach Visualisierungsmöglichkeiten umfangreicher gartenbaulicher Datensätze, die einerseits zu einer graphischen Zusammenfassung der in den Daten enthaltenen Informationen dienen und die andererseits auf interaktivem Weg Möglichkeiten der graphischen Analyse von Erhebungsdaten liefern.

Die weitgehende Freiheit von Modellannahmen, der überwiegend deskriptive Charakter der Untersuchungen, das interaktive, schrittweise Vorgehen in der Auswertung, und die Betonung graphischer Elemente kennzeichnet die Arbeit als Beitrag zur explorativen Datenanalyse.

Das ausgewählte Methodenspektrum, das ausführlich besprochen wird, schließt Verfahren der Dimensionserniedrigung (Hauptkomponentenanalyse, Korrespondenzanalyse und mehrdimensionale Skalierung) und darauf aufbauende Biplots, die Analyse gruppierter Daten (Prokrustes-Rotation und Gruppenanalysemodelle in der Hauptkomponentenanalyse), Linienverbände (Liniendiagramme der formalen Begriffsanalyse, Baumdiagramme und graphische Modelle), sowie ergänzende graphische Verfahren, wie zum Beispiel Trellis-Displays, ein.

Beispielhaft werden eine betriebsbegleitende Untersuchung mit Cyclamen aus der Beratungspraxis der Landwirtschaftskammer Westfalen-Lippe und die Kennzahlen der Jahre 1992 bis 1994 der Topfpflanzenbetriebe des Arbeitskreises für Betriebswirtschaft im Gartenbau aus Hannover analysiert. Neben einer Vielzahl informativer Einzelergebnisse, zeigt die Arbeit auch auf, daß die qualitativ relativ schlechten Datengrundlagen nur selten eindeutige Schlußfolgerungen zulassen. Sie sensibilisiert also in diesem Bereich für die Problematik, die der explorativen Analyse wenig perfekter Daten innewohnt.

Als besonders sinnvolle Hilfsmittel in der graphischen Analyse erweisen sich Biplots, hierarchische Liniendiagramme und Trellis-Displays. Die Segmentierung einer Vielzahl von Objekten in einzelne Gruppen wird durch Klassifikations- und Regressionsbäume vor allem unter dem Gesichtspunkt der Visualisierung gut gelöst, da den entstehenden Baumstrukturen auch die die Segmente bestimmenden Variablen visuell entnommen werden können. Diskrete graphische Modelle bieten schließlich einen guten Ansatzpunkt zur Analyse von multivariaten Beziehungszusammenhängen.

Einzelne, nicht in der statistischen Standardsoftware vorhandene Prozeduren sind in eigens erstellten Programmcodes zusammengefaßt und können mit dem Programm Genstat genutzt werden.

LEBENSLAUF

Persönliche Daten

Name:	Stefan Krusche
Geburtstag:	17.05.1959
Geburtsort:	Berlin
Nationalität:	deutsch
Familienstand:	verheiratet, zwei Kinder
Fremdsprachen:	Englisch, Spanisch
EDV-Kenntnisse:	Statistik (Genstat, SPSS, S-Plus)
	Anbauplanung (Gartplan, Greenhouse Care System)
	Standardsoftware (Office Programme)
Führerschein:	Klasse 3

Ausbildung

Studien:

Hochschule
Oktober 1989 - September 1990
University of Reading (England)
Department of Agriculture and Food
Tropical Agricultural Development (Crops)
Verleihung des Titels Master of Science (MSc)
mit Auszeichnung am 15.12.1990

Fachhochschule
September 1985 - September 1989
Fachhochschule Osnabrück, Fachbereich Gartenbau
Vertiefungsfächer: Betriebslehre, Bodenkunde
Versuchswesen, Zierpflanzenbau
Diplomprüfung bestanden am 11.9.1989, Note "sehr gut"

Meisterschule
September 1984 - August 1985
Fachschule für Gartenbau und Weinbau, Veitshöchheim
Gärtnermeisterprüfung bestanden am 31.7.1985, Note "gut"

Berufsausbildung:	1978 - 1980	Ausbildung zum Gärtner (Zierpflanzenbau) bei Willi Valerius & Söhne, Berlin
Schulbildung:	1970 - 1978	Humanistisches Gymnasium
	1966 - 1970	Grundschule

Berufspraxis

... als Diplom-Ingenieur Gartenbau

seit Oktober 1992
Landwirtschaftskammer Westfalen-Lippe
Arbeitsbereichsleiter Zierpflanzenbau/Informationssysteme
im Gartenbauzentrum Westfalen-Lippe - Gartenbauberatung -

Juli 1991 - Juni 1992
Pöppelmann GmbH & Co., Lohne
Berater für TEKU in Spanien

... in Kuzzeiteinsätzen

Mai 1993
Produkt und Markt, Wallenhorst
Untersuchung des marrokanischen Schnittblumenanbaus

November 1988 - Januar 1989
CAAP, Cayambe, Ecuador
Versuchswesen, Kulturen des andinen Hochlandes

... als Gärtnergehilfe

Mai 1983 - April 1984
Harald Weißmann, Würzburg
Zierpflanzen, Gemüse, Endverkauf

März 1981 - Januar 1983
Friedrich Wieland, Berlin
Topfpflanzen, Stauden

September 1980 - Januar 1981
Willi Valerius & Söhne, Berlin
Orchideen, Schnittblumen

Veröffentlichungen (Beispiele)

Entscheidung auf solider Datenbasis	Gärtnerbörse 23/97, Seite 1266 - 1269
Graphical Tracking	Gärtnerbörse 8/97, Seite 438 - 440
Versuchsauswertung per EDV	Gärtnerbörse 29/95, Seite 1272 - 1276

Vorträge (Beispiele)

(K)eine Information zu den Produktionskosten von Cyclamen	Cyclamenseminar, BVG Wolbeck, 18.2.1998
Exploratory analysis of multivariate horticultural data	International Genstat Conference, Wellesbourne, 2.9. - 5.9.1997
Parameterfreie Analyse ordinalskalierter Variablen aus faktoriellen Versuchen	SPSS Anwendertagung München, 15.10 - 16.10.1996

Danksagung

„...., the trajectories reflect the reality and suggest that what looks simple and straightforward may give a very misleading picture of the reality."

John Gower, 30.12.1997

Ich möchte danken:
Prof. Dr. E. Thomas für die Überlassung des Themas und die Betreuung während der Dissertation
Dr. Rudolf Blum für seine Ermutigung die Dissertation in Angriff zu nehmen und fortzuführen
Herrn Hartmuth Range für seine kollegiale Unterstützung bei der Realisierung des Projekts
Herrn Frank Gehring für Unterstützung im Bereich der Bereitstellung entsprechender Hardware
Dr. Ralf Uhte und Herrn Alfons Jokiel für die Bereitstellung der Beispielsdaten

Und all denen, die durch die Entwicklung entsprechender Software diese Arbeit überhaupt erst möglich gemacht haben, und zwar vor allem:
Numerical Algorithms Group, Oxford (Genstat)
StatSci, Oxford (S-Plus)
SPSS Deutschland, München (SPSS, BMDP Diamond, Amos)
Ernst Schröder Zentrum, Darmstadt (Toscana und Anaconda für die formale Begriffsanalyse)
Svend Kreiner, Kopenhagen (Digram zur Auswertung diskreter graphischer Modelle)
Lenny Ravitz, Lexington (PrintGL)
Visual Numerics, Houston (Stanford Graphics)

Erklärung

Ich versichere, daß ich diese Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt habe. Die Stellen, die dem Wortlaut oder Sinn nach anderen Werken entnommen worden sind, habe ich unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt worden.

Münster, den 19.2.2000, Stefan Krusche

Inhalt

1 Einführung

1.1 Motivation, Zielsetzung und Vorgehensweise

Die vorliegende Arbeit befaßt sich mit der Visualisierung, also der graphischen Abbildung, und der Analyse multivariater Daten. Sie ist motiviert durch die Vielzahl an multivariaten Datensätzen, mit denen die Beratung im Gartenbau, in der der Verfasser seit 1992 tätig ist, konfrontiert wird. Die verwendeten Daten, die mit den Methoden visualisiert und analysiert werden, stammen im weitesten Sinne aus der gartenbaulichen Beratung. Multivariate Datensätze tauchen unter anderem im gartenbaulichen Versuchswesen (Sortenversuche), im Einzelhandelsgartenbau-Marketing (Kundenstrukturanalysen), in der Düngerberatung (Substratanalysen), in der Betriebswirtschaft (Kennzahlenvergleiche) oder auch in so beratungstypischen Feldern wie betriebsbegleitenden Untersuchungen auf. Die Qualität der Daten, zum Beispiel im Hinblick auf Repräsentativität, Vollständigkeit oder Vorliegen der Multinormalverteilung, ist allerdings häufig gering. Aus diesem Grund beschränkt sich die Arbeit überwiegend auf beschreibende, nicht schließende Verfahren und ein exploratorisches Vorgehen (siehe 1.2).

Zum Einsatz kommen verschiedene, noch recht junge datenanalytische Verfahren, die zur Visualisierung eingesetzt werden können. Ziel der Arbeit ist es, diese weniger bekannten, datenanalytischen Verfahren für den Gartenbau zu erschließen und sowohl durch ihren Einsatz einen vertieften Einblick in die vorhandenen, gartenbaulich relevanten Daten zu gewinnen, als auch auf Grundlage der gewonnen Erfahrungen eine Beurteilung der Verfahren vornehmen zu können.

Zur Illustration der vorgestellten Methoden dienen:

1. eine betriebsbegleitende Untersuchung der Cyclamenkultur in 20 westfälisch-lippischen Gartenbaubetrieben von 1994 (JOKIEL & HOCKWIEN, 1994) und
2. die Kennzahlen des Kennzahlenvergleichs für Topfpflanzenbetriebe des gesamten Bundesgebietes der Jahre 1992 bis 1994 des Arbeitskreises Betriebswirtschaft in Hannover (UHTE, 1997).

Um die in diesen Daten enthaltenen Informationen zusammenzufassen und darzustellen und um die wesentlichen Aspekte von den unwesentlichen Aspekten zu trennen, gibt es eine Vielzahl von Verfahren, die in ihrer Mehrzahl der multivariaten Statistik zugerechnet werden. Mit ihrer Hilfe gelingt es, graphische Abbildungen zu entwickeln, die die unüberschaubare Datenfülle erfaßbar und interpretierbar macht.

Die Analyse multivariater Daten stellt ein sehr großes Teilgebiet der (statistischen) Datenanalyse dar. Einen umfassenden Überblick über die Vielzahl der Methoden geben KRZANOWSKI & MARRIOTT, 1994 und 1995. Neue Entwicklungen, vor allem auf dem Gebiet der beschreibenden, multivariaten Statistik, werden in KRZANOWSKI, 1995, diskutiert. Vor allem graphische Verfahren der Datenanalyse werden in NAGEL et al., 1996, besprochen. Die für diese Arbeit ausgewählten Methoden stellen nur einen Ausschnitt aus dem Bereich entsprechender Verfahren dar. Bei der

Methodenwahl wird vor allem Wert darauf gelegt, daß sich die Methoden zur Visualisierung eignen und somit:

1. (Ähnlichkeits-)Beziehungen zwischen den beobachteten Objekten (Gegenständen) abbilden,
2. Beziehungen zwischen den, an den Objekten bestimmten, Variablen (Merkmalen) aufzeigen, und
3. Objekte und Merkmale gleichzeitig darstellen.

Da darüber hinaus in den vorliegenden Daten sowohl Objekte als auch Variablen in Objekt- beziehungsweise Variablengruppen vorliegen, oder eine derartige Gruppenbildung möglich und sinnvoll ist, widmet sich die Arbeit auch Methoden, die die Analyse gruppierter Daten in den Vordergrund stellen[1]. Schließlich ist der Überprüfung der Modellvoraussetzungen (soweit erforderlich), der Beurteilung der Darstellungsgüte, und der Einschätzung der Stabilität von graphischen Darstellungen Beachtung zu schenken.

Es werden daher besprochen und eingesetzt:

1. Verfahren der Dimensionserniedrigung und darauf aufbauende Biplots;
2. Verfahren zur Analyse gruppierter Daten;
3. Datenanalyse durch Linienverbände (formale Begriffsanalyse, Baumdiagramme, graphische Modellierung); und
4. weitere graphische und ergänzende Verfahren.

Verfahren der Gruppierung und Klassifizierung, wie zum Beispiel Clusteranalyse, Diskriminanzanalyse oder neuronale Netze, deren graphische Elemente eher von untergeordneter Bedeutung sind[2], sowie multivariate Varianz- und Regressionsanalyse, deren konfirmatorische Bestandteile, zum Beispiel Hypothesentests für die Gleichheit von Mittelwertsvektoren unterschiedlicher Behandlungen, Modellannahmen machen, die durch die Beispielsdaten in der Regel verletzt werden, bleiben weitgehend unberücksichtigt.

Die Arbeit hat drei Schwerpunkte:

1. die Vermittlung neuer Erkenntnisse zu bekannten Daten mit Hilfe der genannten datenanalytischen Verfahren;

---

[1] Datensätze, die eine Objekt- oder Variablengruppierung aufweisen werden in der vorliegenden Arbeit als gruppierte Daten bezeichnet.

[2] Natürlich beinhaltet zum Beispiel die Clusteranalyse auch graphische Elemente, die durchaus in der Analyse genutzt werden (siehe 2.5.1.2). Primäres Ziel clusteranalytischer Verfahren ist aber nach BACHER, 1994, Seite 1, „ ... das *Auffinden einer empirischen Klassifikation*“ , und nicht die graphische Abbildung multivariater Daten.

2. die Erschließung und Beurteilung dieser Verfahren hinsichtlich ihrer Stärken und Schwächen, sowohl im Bereich der Gewinnung als auch im Bereich der Kommunikation von in den Daten enthaltenen Informationen; und schließlich,
3. die Erarbeitung von Programmabläufen für interessante, aber in vielen statistischen Programmen nicht enthaltene Verfahren.

Für die exploratorische Datenanalyse im Rahmen dieser Arbeit werden die folgenden Arbeitshypothesen formuliert:

Die exploratorische Datenanalyse führt zu:

1. einer übersichtlichen Darstellung der vorhandenen Informationen;
2. einer umfassenden Formulierung von Wirkungszusammenhängen;
3. der Aufdeckung (verborgener) Strukturen; sowie
4. der Vermeidung voreiliger Schlußfolgerungen.

Für den weiteren Ablauf der Arbeit wird das folgende Vorgehen gewählt: nach einer kurzen Darstellung dessen, was unter exploratorischer Datenanalyse verstanden wird (im folgenden Abschnitt, 1.2), erfolgt zunächst die Betrachtung der Methodik der ausgewählten Verfahren (Kapitel 2); die besprochenen Methoden werden dann zur Analyse der oben genannten Datensätze eingesetzt (Kapitel 3); schließlich werden die Ergebnisse unter inhaltlichen und methodischen Gesichtspunkten diskutiert (Kapitel 4). Der sehr umfangreiche Bestand an Ergebnistabellen und Ergebnisabbildungen, der ein wesentliches Merkmal der Arbeit darstellt, ist separat im Anhang Teil I und Teil II zusammengefaßt, um den Text leichter lesbar zu gestalten. Einzelne methodische Bereiche, die durch die dem Verfasser vorliegende Software nicht zufriedenstellend gelöst werden können, werden durch Genstat-Codes aufgearbeitet und ergänzen somit das bereits in Genstat vorhandene multivariate Methodenspektrum[3]. Sie sind im Anhang Teil III zu finden (einschließlich der entsprechenden Codes und der Originaldaten).

[3] Genstat ist ein in FORTRAN und C++ programmiertes, allgemeines Statistikpaket, das durch seine Flexibilität und Leistungsfähigkeit für die gegebene Arbeit angemessen ist. Die Möglichkeit der Implementierung eigener Befehlsabläufe in Form sogenannter Prozeduren ermöglicht eine sehr effiziente Einbindung neuer Methodologie in bestehende Programmstrukturen. Andere statistische Programme sind in der Lage ähnliches zu leisten (zum Beispiel SAS oder S-Plus) und der Einsatz von Genstat ist vor allem auf eine persönliche Präferenz des Verfassers zurückzuführen.

1.2 Explorative Datenanalyse - Definition und Beispiele

Da sich die vorliegende Arbeit als Beitrag zur explorativen Datenanalyse gartenbaulicher Daten versteht, gilt es nun den Begriff explorative beziehungsweise exploratorische Datenanalyse zu definieren[4]. Als exploratorische Datenanalyse wird in dieser Arbeit eine Vorgehensweise in der Auswertung von Daten verstanden, in der es primär darum geht, die vorhandenen Daten kennenzulernen, sie aufzubereiten und darzustellen und aus den vorhandenen Daten Hypothesen und Fragestellungen zu entwickeln. Dabei wird in der Regel nicht von einem Wahrscheinlichkeitsmodell ausgegangen, obwohl einige der verwendeten Verfahren durchaus, zumindest implizit, bestimmte Modellannahmen machen (zum Beispiel die kanonische Variablenanalyse, siehe 2.3.3.2). Eine strenge Abgrenzung von Statistik und Datenanalyse, wie sie zum Beispiel von GIFI, 1990, formuliert wird, wird in dieser Arbeit nicht angestrebt. Vielmehr wird der Ansatz von CHATFIELD, 1995, der ein immer stärkeres Ineinanderfließen von explorativer und modellbegründeter, schließender Statistik, für wünschenswert und unvermeidbar hält, vertreten. Die Unterscheidung zwischen EDA (exploratory data analysis) und IDA (initial data analysis), die CHATFIELD, 1995, vornimmt, scheint aber eher künstlich zu sein. Folgt man den für die EDA aufgeführten Methodenkatalogen von BOCK, 1992, oder JAMBU, 1991, und vergleicht diese mit dem im Rahmen der IDA aufgeführten Katalog nach CHATFIELD, 1995, so ergeben sich sehr große Überschneidungen. Zudem ist die Festlegung auf ein bestimmtes Methodenspektrum zur Definition einer datenanalytischen Vorgehensweise ohnehin problematisch, da durch die ständige Weiter- und Neuentwicklung der Methodik, eine ständige Weiterentwicklung der entsprechenden Kataloge erfolgen muß. Wenn zum Beispiel BOROVCNIK, 1992, die Hauptkomponentenanalyse noch als nicht exploratorisch bezeichnet, da ihr die Interaktion zwischen Substanzwissen und mathematischer Darstellung fehlt, so ist dieser Einwand spätestens seit der Weiterentwicklung der Hauptkopmponentenanalyse-Biplots nach GOWER & HAND, 1996, überholt (siehe 2.2). Statt einer Abgrenzung zwischen unterschiedlichen Methodenspektra, wird vielmehr ein stark miteinander verbundenes Analysekonzept entwickelt. Durch die Integration der Computer in die Datenanalyse und die dadurch ständig wachsenden Möglichkeiten, vor allem in der graphischen Datenanalyse, bietet sich eine Durchmischung traditioneller, modellbegründeter, statistischer Verfahren, die von BOCK, 1992, unter dem Begriff CDA (confirmatory data analysis) zusammengefaßt werden, mit parameterfreien, graphischen, und rein deskriptiven Ansatzpunkten in der Auswertung von Daten an[5].

---

[4] Exploratorisch und explorativ werden synonym gebraucht.

[5] Die vorliegende Arbeit verzichtet allerdings fast vollständig auf die Anwendung schließender, statistischer Verfahren, wie zum Beispiel Signifikanztests (bis auf Ausnahmen, zum Beispiel in der Konstruktion von graphischen Modellen oder Klassifikations- und Regressionsbäumen, siehe 2.4.2 und 2.4.3), da die nicht nach auswertungstechnischen Gesichtspunkten gewonnenen Daten (keine Zufallsstichproben, keine Repräsentativität), nicht zu verallgemeinerten Schlüssen über hypothetische Populationen herangezogen werden sollen. Insofern bleiben selbst da, wo

Wie nahe sich explorative und konfirmatorische datenanalytische Ansätze sind, mag die Beschreibung der explorativen Datenanalyse (EDA) nach BOROVCNIK, 1992, verdeutlichen. Als wichtige Merkmale der EDA werden genannt: der Verzicht auf eine Trennung zwischen Theorie und Realität, weitgehende Freiheit von Annahmen, schlechte Eignung für arbeitsteilige Analyse, sowie Probleme bei automatischer Datenanalyse mit nachfolgender Interpretation. Mit Ausnahme der Freiheit von Modellannahmen, die aber eher die Eigenschaft einzelner Verfahren und weniger das Erkennungsmerkmal eines datenanalytischen Ansatzes ist, sind doch für eine Datenanalyse, mit welchen Werkzeugen auch immer, eine automatische Analyse ohne Interaktion zwischen Anwender und Datenanalytiker, oder eine Trennung von Theorie und Realität, die ja nur möglich ist, wenn bewußt an der Realität vorbeigedacht wird, kein wünschenswertes Vorgehen. Ob die explorative Datenanalyse tatsächlich durch eine andere Haltung zur Auswertung von Daten geprägt ist als die konfirmatorische Analyse, wie WOLF, 1992, es darzustellen versucht, mag ebenfalls bezweifelt werden. Das Modell der fünf Explorationsschritte in WOLF, 1992 (Seite 322), entspricht fast vollständig dem Vorschlag zum sinnvollen statistischen Arbeiten, den CHATFIELD, 1995 (Seite 7 und 8), formuliert.

Eine Aufwertung „schlampiger" Analysen im Nachhinein durch Bezeichnung als explorative Studie ist nach BOCK, 1992, Seite 32, eine berechtigte Befürchtung. Darüber hinaus treten als besondere Probleme exploratorischer Analysen auf (BOCK, 1992): Überanpassung, Methodenartefakte, artifizielle interne Widersprüche, Übersehen komplizierter Zusammenhänge, Unschärfe der Begriffswahl und geringe Qualität der Daten, wobei die Frage berechtigt ist, ob dies wirklich ausschließlich und speziell Probleme explorativer Datenanalysen sind. Ebenso sollte die Forderung nach einer „Ethik exploratorischer Datenanalyse" (BOCK, 1992, Seite 31) nicht auf die explorative Datenanalyse beschränkt bleiben (oder geht BOCK, 1992, davon aus, daß diese Ethik im Bereich der CDA per se gegeben ist?).

Wenn sich das Konzept der hier vorgelegten Arbeit im wesentlichen im Bereich dessen bewegt, was gemeinhin als exploratorische Datenanalyse beschrieben wird, so ist diese Vorgehensweise in erster Linie durch die Herkunft der Daten vorgegeben, die eine andere Art der Analyse gar nicht in Betracht kommen läßt. Die Arbeit ist somit 'data driven' (SPRENT, 1997), sowohl in ihrer Konzeptionierung als auch in ihrer Durchführung. Andere Datenqualitäten mögen zu anderen datenanalytischen Vorgehensweisen führen und auch sehr viel konkretere Fragestellungen zu beantworten suchen. Insofern stellt sich für diese Arbeit nicht die Frage, ob die Analyse explorativ oder konfirmatorisch erfolgen soll, sondern vielmehr, ob sie explorativ und/oder konfirmatorisch erfolgen kann.

---

methodenbedingt Signifikanztests verwendet werden, diese ein exploratives Hilfsmittel. Das heißt aber nicht, daß nicht auch im Rahmen der explorativen Datenanalyse, so wie sie in dieser Arbeit verstanden wird, konfirmatorische Verfahren ihren Platz haben können, wenn die Annahmen für ihre Anwendung gegeben sind.

In dieser Arbeit wird der Standpunkt vertreten, daß die Grenzen zwischen unterschiedlichen datenanalytischen Ansätzen fließend sind, daß es in jeder Datenanalyse darauf ankommt, die in den Daten enthaltenen Informationen zu entdecken, zu beschreiben und darzustellen, und daß die Auswertungsmethodik vor allem durch Herkunft, Qualität und Struktur des Datenmaterials, sowie durch die zu beantwortenden Fragestellungen bestimmt wird. Es ist allerdings zu klären, ob die Vielzahl der verfügbaren datenanalytischen Verfahren, in der Lage ist, wichtige Informationen zu liefern, oder ob sie nicht auch zu Beliebigkeit und zunehmender Unklarheit der Ergebnisse führt. In diesem Zusammenhang ist darauf hinzuweisen, daß die Arbeit auch Merkmale aufweist, die der modellfreien Analyse derartig wenig 'perfekter' Daten inhärent sind, das heißt, es ist nicht zu erwarten, daß unterschiedliche Analyseschritte zu ausschließlich widerspruchsfreien Ergebnissen und immer eindeutig 'richtigen' Interpretationen führen. Sie ist in gewissem Sinne auch ein 'data mining', das auf den verschiedensten Wegen versucht, interessante Aspekte in den Daten zu entdecken, ohne einen Allgemeingültigkeitsanspruch geltend machen zu können[6].

Als Schlußfolgerung ist für diese Arbeit und generell für das Vorgehen in der (explorativen) Datenanalyse festzuhalten:

1. es ist zu formulieren, was durch den Einsatz eines bestimmten Verfahrens gezeigt werden soll und kann;
2. es ist zu begründen, warum ein Vorgehen gewählt wird, beziehungsweise offenzulegen, worauf die Auswahl eines Verfahrens beruht; und
3. es ist klarzustellen, wo die Grenzen der angewendeten Methodik und der durch sie gewonnenen Ergebnisse liegen.

Um diesen Anforderungen gerecht zu werden, ist dem Kapitel 3 (der Ergebnisdarstellung), das Kapitel 2 mit der Erläuterung der methodischen Grundlagen der verwendeten Verfahren vorangestellt.

Abschließend sollen einige Arbeiten aus dem Gartenbau erwähnt werden, die sich multivariater Verfahren bedienen und im weitesten Sinne als explorativ verstanden werden können. Die Diskussion um Sinn oder Unsinn multivariater Verfahren in Gartenbau und Landwirtschaft ist nicht neu und wird zum Beispiel bereits von FINNEY, 1956, oder PEARCE & HOLLAND, 1960, kontrovers geführt. FINNEY, 1956, folgert bezüglich des Einsatzes von multivariater Varianzanalyse und kanonischer Variablenanalyse (unter Bezugnahme auf eine Arbeit von STEEL, 1955):

„ ... in field experiments and in many other research problems the type of multivariate analysis illustrated by Steel is usually inappropriate and often actively misleading." (Seite 71).

---

[6] Allerdings wird der Begriff 'data mining' in der Regel im Zusammenhang mit noch sehr viel größeren Datensätzen als sie hier vorliegen, verwendet (CHATFIELD, 1997).

Demgegenüber äußern PEARCE & HOLLAND, 1960, mit Blick auf Hauptkomponenten- und Faktoranalyse in der Auswertung von Gehölzdaten die Meinung: „It does appear, therefore, that multivariate methods can form a useful extension of the more accepted methods and can lead to a better understanding of the tree as a whole." (Seite 7). Unabhängig von diesen sehr unterschiedlichen Bewertungen ist festzustellen, daß multivariate Verfahren in vielen Bereichen Eingang in die gartenbauliche Forschung gefunden haben. Als Beispiele seien genannt: die Arbeiten von BAUER & TEUTTER, 1992, BENNE, 1990, CRUZ-CASTILLO et al., 1994, FISCHER, 1993, RATH, 1996, und STEINBACHER et al., 1995, im Bereich der Klassifikationsverfahren (Diskriminanzanalyse und neuronale Netze) zur Unterscheidung von zum Beispiel Sorten oder Wachstums- und Entwicklungsstadien; die Arbeiten von BEYL et al.,1995, DEGANI et al., 1995, FABBRI et al., 1995, NOVI et al., 1996, PEREIRA-LORENZO et al, 1996a und 1996b, und REN et al., 1995, die die Clusteranalyse und Dendrogramme auf dem Gebiet der Pflanzenzüchtung einsetzen; die Arbeiten von DEVER et al., 1996, FERNANDEZ et al., 1996, NIENHUIS et al., 1996, PARENT et al, 1994, PLOTTO et al., 1997, RUMAYOR-RODRÍGEZ, 1995, und TIVANG et al., 1996, in denen Verfahren der Dimensionserniedrigung und rudimentär auch Biplots, eingesetzt werden, um Nährstoffgehalte zu analysieren, unterschiedliche Genotypen zu identifizieren oder um die Ergebnisse von Geschmacksanalysen sichtbar zu machen. Umfassendere Studien, die mit einer Vielzahl von Verfahren arbeiten, und damit den explorativen Ansatz ihrer Arbeiten deutlich machen, sind zum Beispiel THOMAS, 1992, oder BARÁTH, 1993. Weitere Beispiele, die einen konkreten inhaltlichen Bezug zu den Auswertungen in dieser Arbeit haben, werden in Kapitel 3 angesprochen.

2 Erläuterung des Methodenspektrums

Da die Darstellung der verwendeten Methoden nicht ohne mathematische Formeln auskommt, ist an dieser Stelle auf einige Konventionen hinzuweisen. In Formeln stehen fettgedruckte Großbuchstaben für Matrizen, fettgedruckte Kleinbuchstaben für Zeilenvektoren. Zeilen- und Spaltenanzahl einer Matrix werden in Klammern nach dem Muster: (Anzahl Zeilen x Anzahl Spalten) angegeben. Das „ ' “-Zeichen steht für die Transposition einer Matrix beziehungsweise eines Vektors. Die Benennungen in den verwendeten Formeln sind so gewählt, daß eine größtmögliche Eindeutigkeit besteht, obwohl bisweilen Überschneidungen vorkommen, die aber dann im entsprechenden Kontext erklärt werden. Nach STEVENS, 1951, wird in dieser Arbeit die Einteilung von Variablen in nominal, ordinal-, intervall- und verhältnisskalierte Variablen verwendet.

2.1 Verfahren der Dimensionserniedrigung

2.1.1 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse ist eine Analysetechnik, mit deren Hilfe p korrelierte Variablen, die an n Objekten (i = 1 ... n) bestimmt werden, in p (j = 1 ... p) neue, nicht korrelierte Variablen, die sogenannten Hauptkomponenten, transformiert werden. Die Transformation ist so gewählt, daß die erste Hauptkomponente den größten Anteil der Gesamtvariabilität der Ausgangsvariablen repräsentiert, die zweite Hauptkomponente den zweitgrößten Anteil, die dritte Hauptkomponente den drittgrößten Anteil und so weiter. Im günstigsten Fall reichen für die Beschreibung der Variabilitätsstruktur der Ausgangsvariablen einige wenige Hauptkomponenten aus, so daß durch die Hauptkomponentenanalyse eine wesentliche Dimensionserniedrigung möglich ist.

Die Hauptkomponentenanalyse ist eine variablenbezogene Methode und zählt zu den R-Techniken, das heißt Ausgangspunkt der Analyse ist eine (p x p) Kovarianz- oder Korrelationsmatrix. Sie wird für die Analyse intervall- beziehungsweise verhältnisskalierter und ordinalskalierter Variablen eingesetzt. Die Verwendung von zum Beispiel dichotomisierten, nominalskalierten Variablen ist nach KRZANOWSKI, 1988a, möglich, kann aber zu stark ausgeweiteten Datensätzen und zu erheblichen Schwierigkeiten bei der Interpretation der Ergebnisse führen.

Ihre wesentliche Bedeutung hat die Hauptkomponentenanalyse als parameterfreie, deskriptive Methode, obwohl auch Elemente der schließenden Statistik eingebracht werden können, sofern die Annahme getroffen werden kann, daß die untersuchten Objekte eine Stichprobe aus einer multinormalverteilten Grundgesamtheit darstellen[7]. Da aber

[7] Im wesentlichen Hypothesentests (zum Beispiel Test auf Gleichheit eines Eigenvektors der Stichprobe mit einem hypothetischen Eigenvektor der Grundgesamtheit) und die Berechnung von Vertrauensintervallen (zum Beispiel für den größten Eigenwert); siehe zum Beispiel ANDERSON, 1963, BARTLETT, 1950 & 1954, LAWLEY, 1956, SCHOTT, 1988.

1. die Annahme der Multinormalverteilung relativ häufig nicht zutrifft,
2. Stichprobe und Grundgesamtheit bisweilen identisch sind, und
3. viele der für die Hauptkomponentenanalyse entwickelten schließenden Verfahren nur asymptotisch gelten (CHATFIELD & COLLINS, 1980),

ist die Bedeutung der Hauptkomponentenanalyse als beschreibende Methode ohne ein der Analyse zugrunde liegendes, statistisches Modell größer als ihre Bedeutung im Bereich der konfirmatorischen Statistik. In der vorliegenden Untersuchung steht - beim Einsatz der Hauptkomponentenanalyse - die durch sie zu erzielende Dimensionserniedrigung und Erklärung der Variablenstruktur im Vordergrund. Darüber hinaus spielt auch die Verwendung (ausgewählter) Hauptkomponenten an Stelle der Ausgangsvariablen in Folgeanalysen, vor allem in der graphischen Repräsentation der Objekte, eine Rolle. Ausführlich wird die Hauptkomponentenanalyse in einer Vielzahl von Standardwerken zur multivariaten Statistik behandelt (siehe unter anderem MORRISON, 1990, oder JACKSON, 1991, und die darin genannten Quellen).

Ausgangspunkt für die Berechnung der Hauptkomponenten ist die (p x p) Kovarianzmatrix **S** der (n x p) Datenmatrix **X**. Im Sinne der Skalierung entspricht die Verwendung der Kovarianzmatrix einer Mittelwertszentrierung der Ausgangsvariablen. Ein wesentlicher Vorteil der Verwendung der Kovarianzmatrix in der Hauptkomponentenanalyse ist die Tatsache, daß die schließenden Verfahren hier relativ gut entwickelt sind. Nachteilig ist es aber, daß, wenn die Variablen in voneinander abweichenden Einheiten bestimmt werden, beziehungsweise stark voneinander abweichende Varianzen aufweisen - eine Situation wie sie im Bereich der gartenbaulichen Daten eher die Regel als die Ausnahme ist - die Variablen mit der größeren Varianz einen ungleich stärkeren Einfluß auf die erste Hauptkomponente ausüben, als die Variablen mit der kleineren Varianz. Um allen Variablen unter solchen Umständen das gleiche Gewicht zu verleihen, werden die Ausgangsvariablen standardisiert, womit an die Stelle von **S** die Korrelationsmatrix **R** tritt. Die Ausgangsvariablen werden also so skaliert, daß ihr Mittelwert = 0 und ihre Varianz = 1 sind. Bei Verwendung von **S** kommt man in der Regel zu anderen Ergebnissen als bei Verwendung von **R**, und keine einfache Transformation kann die Ergebnisse einer auf **S** basierenden Hauptkomponentenanalyse in eine auf **R** basierende Hauptkomponentenanalyse umwandeln. Die Hauptkomponenten sind also kein einzigartiges Merkmal der Ausgangsmatrix, sondern abhängig von der Skalierung der Variablen (KRZANOWSKI, 1988a).

Wenn es gelingt die Datenmatrix **X** statt durch p Ausgangsvariablen mit q ($q < p$) Hauptkomponenten zu beschreiben, ohne daß damit ein nennenswerter Informationsverlust einhergeht, wird durch die Hauptkomponentenanalyse eine (wünschenswerte) Dimensionserniedrigung von **X** erreicht. Besonders vorteilhaft ist es, wenn $q = 2$ ist, da dann zweidimensionale Graphiken, in denen zum Beispiel die Hauptkomponentenwerte der zweiten Hauptkomponente gegen die Hauptkomponentenwerte der ersten Hauptkomponente geplottet

werden, einen guten Einblick in die Struktur und die Beziehungen der Objekte untereinander ermöglichen. Vorhandene Gruppierungen der Objekte können so möglicherweise erkannt werden.

Der Frage, welche Hauptkomponenten näher betrachtet werden sollten, welche Hauptkomponenten also 'wichtige' oder 'wesentliche' Hauptkomponenten darstellen, sind zahlreiche Arbeiten nachgegangen. Es geht ihnen allen darum, ein Kriterium festzulegen, das die Entscheidung unterstützt, welche Hauptkomponenten berücksichtigt werden sollen, welche Hauptkomponenten also ins reduzierte Modell zur Beschreibung von **X** aufgenommen werden sollen, und welche Hauptkomponenten verworfen werden können. In Tabelle 1 sind aus diesen Arbeiten einige der gebräuchlichsten Kriterien zur Identifikation 'wesentlicher' Hauptkomponenten zusammengefaßt.

Zusätzlich ist noch auf zwei weitere Verfahren, die in Kapitel 3 verschiedentlich eingesetzt werden, hinzuweisen, und zwar auf die partielle Korrelations-Prozedur nach VELICER, 1976, und die Variante der Kreuzvalidierung nach EASTMENT & KRZANOWSKI, 1982. *Für beide liegt der entsprechende Genstat Code im Anhang Teil III vor.*

Die 1976 von VELICER vorgeschlagene Methode zur Identifikation 'wesentlicher' Hauptkomponenten verwendet als Entscheidungskriterium die partielle Korrelation zwischen den Ausgangsvariablen, für den Fall, daß q Hauptkomponenten (q = 0, ... , p - 1) aus dem Modell entfernt werden (das heißt keine Hauptkomponente, die erste Hauptkomponente, die erste und die zweite Hauptkomponente, die erste, die zweite und die dritte Hauptkomponente und so weiter). Der zu errechnende Wert f_q bei Entfernung von q Hauptkomponenten, der im wesentlichen durch die die Quadratsumme der partiellen Korrelationen bestimmt wird, hat ein Minimum im Bereich von $0 < q < p - 1$. Der Wert von q, bei dem das Minimum von f_q erreicht wird, gibt die Anzahl der q 'wesentlichen' Hauptkomponenten an. Solange f_q abnimmt, nehmen die partiellen Korrelationen stärker ab als die Restvarianzen, das heißt die Varianzen der nicht im Modell berücksichtigten Hauptkomponenten.

Die Anwendung der Kreuzvalidierung zur Identifikation 'wesentlicher' Hauptkomponenten geht zurück auf die Arbeit von EASTMENT & KRZANOWSKI, 1982. Kreuzvalidierung ist sowohl bei Verwendung der Kovarianz- als auch der Korrelationsmatrix in der Hauptkomponentenanalyse möglich. Im Prinzip geht es um folgendes: von einer Anzahl m von Modellen mit den Parametern $\beta^{(m)}$, soll eines ausgewählt werden. Die Datenmatrix **X** mit n Objekten soll die Modellwahl bestimmen. Wenn nun das i-te Objekt von **X** gelöscht wird, können auf Grundlage der n - 1 Objekte von **X** die Parameter $\hat{\beta}^{(m)}$, und mit $\hat{\beta}^{(m)}$ die Werte des gelöschten Objekts $\mathbf{x_i}$ als $\hat{\mathbf{x}}_\mathbf{i}^{(m)}$ geschätzt werden. Aus der Abweichung der beobachteten Werte $\mathbf{x_i}$ von den geschätzten Werten $\hat{\mathbf{x}}_\mathbf{i}^{(m)}$, läßt sich ein Diskrepanzmaß bestimmen, allgemein $f(\hat{\mathbf{x}}_\mathbf{i}^{(m)}, \mathbf{x_i})$. Wenn dieser Vorgang für alle Objekte von **X** wiederholt wird, ergibt sich ein Diskrepanzmaß $E(m) = k\sum_{i=1}^{n} f(\hat{\mathbf{x}}_\mathbf{i}^{(m)}, \mathbf{x_i})$ E(m) für Modell m mit Korrekturfaktor k. Ein Vergleich der Diskrepanzmaße, von m Modellen gibt Aufschluß darüber,

mit welchem Modell die kleinste Abweichung von geschätzten und beobachteten Werten erreicht wird. Das Diskrepanzmaß im Verfahren von EASTMENT & KRZANOWSKI, 1982, ist die sogenannte PRESS (Prediction Sum of Squares)-Statistik:

$\mathrm{PRESS}^{(m)} = 1/\mathrm{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\hat{x}_{ij}^{(m)} - x_{ij}\right)^2$. Der Vergleich der Diskrepanzmaße errechnet sich durch

$\mathrm{W}^{(m)} = \left[\left(\mathrm{PRESS}^{(m-1)} - \mathrm{PRESS}^{(m)}\right)/\left(n + p - 2m\right)\right]/\left[\mathrm{PRESS}^{(m)}/p(n-1)\right]$, wobei m für die Anzahl der im Modell betrachteten Hauptkomponenten steht. Der W-Wert, kleiner 1, der dem Modell mit der geringsten Anzahl an Hauptkomponenten entspricht, dient als Hinweis auf die Anzahl ‘wesentlicher’ Hauptkomponenten.

Unterschiedliche Kriterien zur Bestimmung der Anzahl der ‘wesentlichen’ Hauptkomponenten führen in der Regel auch zu unterschiedlichen Schlußfolgerungen. Insofern sind die genannten Verfahren nur als Anhaltspunkte für die Anzahl der im Modell zu berücksichtigenden Hauptkomponenten zu verstehen. Neben dem Skalierungsproblem ist die Tatsache, daß unterschiedliche Kriterien häufig zu unterschiedlichen Schlußfolgerungen bezüglich der Auswahl 'wesentlicher' Hauptkomponenten führen, für Kritiker der Hauptkomponentenanalyse wie CHATFIELD & COLLINS, 1980, eines der stärksten Argumente, mit der sie die Hauptkomponentenanalyse kritisieren.

Unter den Residuen der Hauptkomponentenanalyse werden in dieser Arbeit die Abweichungen der durch das Hauptkomponentenmodell reproduzierten Werte der Ausgangsvariablen von den tatsächlichen Beobachtungs- oder Meßwerten verstanden. Eine Analyse der Residuen kann zeigen, in wie weit das gewählte, erniedrigte Modell zu den Beobachtungswerten paßt. Objekte mit einem sehr großen Residuum werden durch das gewählte Modell schlecht repräsentiert. Große Residuen können die Folge von tatsächlich stark von den übrigen Objekten abweichenden Beobachtungen oder aber auch von Aufzeichnungs- und Übertragungsfehlern sein. Teststatistiken für die Residuen und kritische Werte geben zum Beispiel HAWKINS, 1974 und 1980, sowie JACKSON, 1991. Nach JACKSON, 1991, ist $Q_\alpha = \theta_1\left[\left(c_\alpha\sqrt{2\theta_2 h_0^2}/\theta_1\right) + \left(\theta_2 h_0\left(h_0 - 1\right)/\theta_1^2\right) + 1\right]^{1/h_0}$ ein kritischer Wert, mit $\theta_1 = \sum_{j=q+1}^{p} l_j$, $\theta_2 = \sum_{j=q+1}^{p} l_j^2$, $\theta_3 = \sum_{j=q+1}^{p} l_j^3$, $h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$, l_j als dem Eigenwert der j-ten Hauptkomponente und c_α als dem Wert der Funktion der Standardnormalverteilung bei Irrtumswahrscheinlichkeit α mit demselben Vorzeichen wie h_0. Wird dieser vom Residuum eines Objekts überschritten, so ist dies ein Indiz dafür, daß das Objekt mit dem entsprechend hohen Residuum nicht adäquat durch die gewählte Dimensionserniedrigung repräsentiert wird. *Genstat Codes zur Erzeugung von Residuenplots und der Berechnung der entsprechenden Statistiken sind im Anhang Teil III zu finden.*

Bisweilen wird der Versuch unternommen, den Hauptkomponenten eine bestimmte Interpretation zu geben. Diese Interpretation orientiert sich am Vorzeichen und der Größe der Koeffizienten der

Eigenvektoren. Im günstigsten Fall ermöglicht die Koeffizienteninterpretation eine zusammenfassende Beschreibung mehrerer Variablen mit einem Begriff, so daß, bei Auswahl von wenigen, gut interpretierbaren Hauptkomponenten, die Variablen- und Variabilitätsstruktur umfangreicher Datensätze knapp und prägnant benannt werden kann. Beispiele solcher Interpretationsansätze, sind zum Beispiel bei MANLY, 1986, oder MORRISON, 1990, zu finden. Allerdings ist die Interpretation der Hauptkomponenten häufig mit großen Schwierigkeiten verbunden. CHATFIELD & COLLINS, 1980, warnen vor einer Überinterpretation der Hauptkomponenten. Auch KRZANOWSKI, 1988a, betont, daß in der praktischen Anwendung der Hauptkomponentenanalyse nur selten der Fall gegeben ist, daß eine klare und eindeutige Interpretation der Koeffizienten möglich ist, und es letztlich von der jeweiligen subjektiven Beurteilung des Anwenders abhängt, welche Koeffizienten als groß oder klein genug angesehen werden, um die Interpretation und Beurteilung der Hauptkomponenten wesentlich mitzubestimmen. In noch stärkerem Umfang als bei der Identifikation der ‘wesentlichen’ Hauptkomponenten ist in der Interpretation der Hauptkomponenten ein willkürliches Element enthalten, das zu einer gewissen Beliebigkeit der Ergebnisse beiträgt. MARRIOTT, 1974, kommt gar zu dem Schluß, daß eine gut interpretierbare und mit einer echten inhaltlichen Bedeutung ausgestattete Hauptkomponente nicht mehr als ein glücklicher Zufall sein kann, da kein Rechenverfahren an sich in der Lage ist, ein im jeweiligen Kontext des Anwendungsgebietes bedeutungsvolles Ergebnis zu produzieren. Da auch relativ geringe Veränderungen bei den Werten der Ausgangsvariablen einen relativ starken Einfluß auf die Koeffizienten der Eigenvektoren und damit auf die Interpretation der Hauptkomponenten haben können, ist die Interpretation mit einem weiteren Unsicherheitsfaktor belastet. Dennoch wird es angebracht sein - im Bewußtsein um die Schwierigkeiten und Begrenzungen der Interpretation - den Versuch zu unternehmen, die die Hauptkomponenten dominierenden Variablen zu bennenen und mögliche Unterschiede und Beziehungen der Koeffizienten und Hauptkomponenten untereinander zu verdeutlichen und somit ansatzweise eine Interpretation durchzuführen. Ein Hilfsmittel, das die Hauptkomponenten besser interpretierbar machen kann, ist die Rotation der Hauptkomponenten. Die Rotation soll eine Vereinfachung der Koeffizientenstruktur herbeiführen. Nicht immer kann eine Rotation eine nennenswerte Vereinfachung der Koeffizientenstruktur bewirken und nicht in jedem Fall ist eine Rotation der Hauptkomponenten sinnvoll. Ein zu beachtendes Merkmal der Rotation ist zudem die Tatsache, daß die Koeffizienten der rotierten Hauptkomponenten nicht unabhängig von der Anzahl der im Modell berücksichtigten Hauptkomponenten sind, das heißt wenn q von p Hauptkomponenten rotiert werden, ergeben sich andere Koeffizienten, als wenn q + 1 derselben p Hauptkomponenten rotiert werden.

Die Rotation kann als orthogonale oder schiefwinklige (oblique) Rotation erfolgen. Bei einer orthogonalen Rotation der Eigenvektoren bleibt die Orthogonalität der Koeffizienten der Eigenvektoren der Hauptkomponenten vor und nach der Rotation erhalten. Die neuen, rotierten Hauptkomponentenwerte sind aber nicht mehr in jedem Fall, wie die ursprünglichen Hauptkomponentenwerte, unkorreliert. Bei der schiefwinkligen Rotation kann ebenfalls die

Unkorreliertheit der Hauptkomponentenwerte und darüber hinaus auch die Orthogonalität der Koeffizienten verloren gehen. Diesen negativen Veränderungen steht (hoffentlich) ein erkennbarer Gewinn in Form einer vereinfachten Koeffizientenstruktur gegenüber. Statistische Software bietet im Rahmen der Hauptkomponentenanalyse oder Faktoranalyse eine Vielzahl orthogonaler und schiefwinkliger Rotationsverfahren an, die auf iterativem Weg ein bestimmtes Optimalitätskriterium zu erreichen suchen und so die neuen, rotierten Komponenten erzeugen (CARROL, 1953, HARMAN, 1974, KAISER, 1959). In dieser Arbeit wird jedoch aufgrund der genannten Schwierigkeiten gänzlich auf den Einsatz von Rotationen im Bereich der Hauptkomponentenanalyse verzichtet.

Tabelle 1: Kriterien zur Identifikation 'wesentlicher' Hauptkomponenten

Kriterium	**Vorgehen, Anmerkungen**	**Literatur**
Anteil der durch die Hauptkom-ponenten 'erklärten' Varianz	Hauptkomponenten werden solange ins Modell aufgenommen, bis ein bestimmter Schwellenwert für die 'erklärte' Varianz überschritten wird, häufig 95 % der Gesamtvariabilität. Die Festlegung des Schwellenwertes ist in der Regel willkürlich und daher nicht unproblematisch.	JACKSON, 1991
Gebrochener Stab	Hauptkomponenten werden solange ins Modell aufgenommen bis der Anteil (in Teilen von 1) ‘erklärter’ Varianz der jeweiligen Hauptkomponente kleiner ist als $g_q = 1/p\sum_{j=q}^{p}(1/j)$. Grundgedanke ist hier, daß Hauptkomponenten solange ins Modell aufgenommen werden, solange die durch sie 'erklärte' Varianz größer ist als die 'erklärte' Varianz ist, die man auch bei rein zufälliger Aufteilung der Gesamtvarianz mit der q-ten Einteilung erklären könnte, das heißt als g_q.	JOLIFFE, 1986
Mittlerer Eigenwert	Hauptkomponenten werden solange ins Modell aufgenommen, solange der Eigenwert der jeweiligen Hauptkomponente größer als der mittlere Eigenwert ist. Ein häufig bei der Verwendung der Korrelationsmatrix eingesetztes Kriterium, da mittlerer Eigenwert von **R** gleich 1, und Varianz der Ausgangsvariablen ebenfalls gleich 1. Liegt der Eigenwert der Hauptkomponente unter 1, so wird durch diese Hauptkomponente weniger Variabilität repräsentiert als durch eine Ausgangsvariable.	GUTTMANN, 1954 JOLIFFE, 1972, JOLIFFE, 1973
Scree-Diagramm	Diagramm mit Eigenwerten (eventuell den Logarithmen der Eigenwerte) auf der y-Achse, der laufenden Nummer des Eigenwertes auf der x-Achse. Hauptkomponenten werden bis zu dem Punkt ins Modell aufgenommen, an dem ein Bruch im Diagramm auftaucht, und die Eigenwerte beginnen, sich sehr ähnlich zu sein. Aufgenommen werden alle Hauptkomponenten bis zur Bruchstelle (inklusive der ersten Hauptkomponente im Verflachungsbereich des Diagramms). Mögliche Probleme: keine deutliche Bruchstelle oder mehrere Bruchstellen. HORN, 1965, schlägt zur Verbesserung bei der Entscheidungsfindung in der Modellauswahl, Generation von Zufallsdaten und Vergleich der Eigenwerte der Zufallsdaten mit den Eigenwerten der Untersuchungsdaten, vor.	CATTEL, 1966
Signifikanztests	Häufig werden durch Signifikanztests sehr viele Hauptkomponenten ins Modell aufgenommen. Grundsätzlich ist die Frage zu klären, ob die Voraussetzungen für die Anwendung der Tests gegeben sind. Nicht alle signifikanten Hauptkomponenten müssen notwendigerweise ins Modell aufgenommen werden. Allerdings sollten - bei Vorliegen der Testvoraussetzungen - nicht Hauptkomponenten aufgenommen werden, die nicht signifikant sind (JACKSON, 1991).	ANDERSON, 1963, LAWLEY, 1956
Tolerierte Restvarianz	Hauptkomponenten werden ins Modell aufgenommen, bis der, vor Beginn der Analyse festgelegte Schwellenwert für die zu tolerierende Restvarianz, noch nicht erreicht ist. Vor allem dort ein sinnvolles Kriterium, wo a priori eine Kenntnis über die inhärente Variabilität der Variablen vorhanden ist.	BOX et al., 1973

2.1.2 Mehrdimensionale Skalierung

Im Vordergrund der mehrdimensionalen Skalierung steht - ähnlich wie bei der Hauptkomponentenanlyse - die Dimensionserniedrigung. Sie wird im wesentlichen deskriptiv eingesetzt und beinhaltet nur wenige konfirmatorische Ansätze[8]. Es handelt sich um eine Q-Technik, das heißt Ausgangspunkt der Analyse ist eine (n x n) Proximitätsmatrix, wobei Proximität sowohl für Ähnlichkeit als auch Unähnlichkeit steht. Die Proximitätsmatrizen können entweder direkt ermittelt oder mit Hilfe eines geeigneten Verfahrens aus nominal-, ordinal- und intervall- oder verhältnisskalierten Variablen hergeleitet werden. Da Proximitätsmatrizen auch für nominalskalierte Variablen erstellt werden können, bieten die Q-Techniken gegenüber den R-Techniken den Vorteil der Handhabbarkeit derartiger Variablen beziehungsweise gemischter Datensätze (GORDON, 1981).

Wichtigstes Ziel der mehrdimensionalen Skalierung ist die graphische Repräsentation der Objekte aufgrund ihrer Proximität, das heißt, daß die Ähnlichkeiten beziehungsweise Unähnlichkeiten zwischen den Objekten, in möglichst wenig Dimensionen und mit möglichst geringem Informationsverlust, so graphisch abgebildet werden sollen, daß die (euklidischen) Distanzen zwischen den Objekten in einer Graphik in etwa den tatsächlichen Proximitäten der Objekte entsprechen (YOUNG, 1987).

Ähnliche Fragestellungen wie bei der Hauptkomponentenanalyse treten auch bei der mehrdimensionalen Skalierung auf; so zum Beispiel im Bereich der Skalierung (Mittelwertszentrierung, Standardisierung der Ausgangsvariablen vor Berechnung von Proximitätsmatrizen), bei der Bestimmung der angemessenen Zahl der zu betrachtenden Dimensionen und der Frage nach ihrer Interpretierbarkeit. Auch das weitgehende Fehlen schließender Verfahren ist zu beachten.

Ausgangspunkt einer Analyse durch mehrdimensionale Skalierung sind eine oder mehrere Proximitätsmatrizen. Die Umwandlung eines Ähnlichkeitsmaßes in ein Unähnlichkeitsmaß (und umgekehrt) ist auf verschiedenen Wegen möglich, zum Beispiel durch: Unähnlichkeitsmaß in Teilen von 1 = 1 - Ähnlichkeitsmaß in Teilen von 1.

Proximitätsmatrizen können entweder direkt ermittelt werden - dies ist für gartenbauliche Daten jedoch die Ausnahme - oder, und das ist der Normalfall, durch ein gewähltes Proximitätsmaß aus den Variablen, die an den jeweiligen Objekten bestimmt werden, hergeleitet werden[9]. Je nachdem,

---

[8] RAMSAY, 1977, 1978, 1980 und 1982 sowie BRADY, 1985, geben einige Anregungen, wie die mehrdimensionale Skalierung durch schließende Verfahren erweitert werden kann. Das von Ramsay entwickelte Programm Multiscale beinhaltet viele dieser Erweiterungen (SCHIFFMAN et al., 1981).

[9] Zu Grundlagen von Proximitätsmaßen siehe zum Beispiel JARDINE & SIBSON, 1972.

ob es sich bei dem Proximitätsmaß um ein Ähnlichkeits- oder Unähnlichkeitsmaß handelt, beziehungsweise die Variablen als nominal-, ordinal- oder intervall- beziehungsweise verhältnisskaliert betrachtet werden können, gibt es eine Vielzahl von Maßen; einen umfassenden Überblick geben zum Beispiel BACHER, 1994 oder SCHUBÖ et al., 1991. Einige Proximitätsmaße sind in Tabelle 2 aufgeführt.

Die Entscheidung für die Verwendung eines bestimmten Proximitätsmaßes wird einerseits durch das Skalenniveau der Variablen bestimmt. Darüber hinaus soll das Proximitätsmaß aber auch die Beziehung zweier Objekte wiederspiegeln können. Ist zum Beispiel das Fehlen eines Merkmals beim Vergleich zweier Objekte unerheblich, die Übereinstimmung aber wichtig, so ist bei binären Variablen dem Jaccard-Ähnlichkeitsmaß der Vorzug vor dem Simple Matching-Ähnlichkeitsmaß zu geben, da in einem solchen Fall das Simple Matching-Ähnlichkeitsmaß die Ähnlichkeit überschätzen würde (durch Überbewertung eines im Grunde irrelevanten Sachverhaltes). Schließlich kann durch die Wahl des Proximitätsmaßes auch Einfluß darauf genommen werden, ob größere oder kleinere Proximitäten mehr Gewicht bekommen sollen, zum Beispiel durch entsprechende Wahl des Exponenten in der Minkowski Metrik. Je höher der Exponent ist, desto größer wird der Unterschied zwischen Objekten mit größerer Unähnlichkeit im Vergleich zu Objekten mit geringerer Unähnlichkeit.

Proximitätsmaße sind in der Regel nicht skalenunabhängig. Insofern ist eine Skalierung in Form einer Mittelwertszentrierung, Standardisierung oder ähnlichem, in Betracht zu ziehen, wenn die gegebenen Daten dies erforderlich erscheinen lassen (KRZANOWSKI, 1988a). Die durch eine Standardisierung erzielte Gleichgewichtung aller Variablen ist in der anfänglichen Phase der Datenanalyse wohl empfehlenswert, keinesfalls aber zwingend (GORDON, 1981). Auch stark korrelierte Variablen können auf das Proximitätsmaß einen (unerwünscht) hohen Einfluß haben. Bisweilen empfohlen, aber nicht unproblematisch, ist dann die Verwendung eines an die Mahalanobis Distanz angelehnten Proximitätsmaßes (DEICHSEL & TRAMPISCH, 1985)[10].

In vielen Fällen liegen Datensätze vor, in denen sich sowohl intervall- beziehungsweise verhältnis-skalierte, als auch nominal- und/oder ordinalskalierte Variablen befinden. Eine Möglichkeit ist dann die Ermittlung separater Proximitätsmatrizen entsprechend des jeweiligen Skalenniveaus und die getrennte Analyse. Eine Alternative ist die Ermittlung einer, aus verschiedenen Proximitätsmaßen gebildeten, mittleren (gewichteten) Proximitätsmatrix. Letzterer Gedanke wird durch den allgemeinen Ähnlichkeitskoeffizienten formalisiert (GOWER, 1971, GOWER & LEGENDRE, 1986).

---

[10] Speziell bei Untersuchungen, die auf Gruppierungen von Objekten hinzielen ist dies problematisch, da das Proximitätsmaß von einer gemeinsamen Kovarianzmatrix für alle Objekte ausgeht, es aber durchaus denkbar ist, daß unterschiedliche Gruppen - die aber a priori nicht bekannt sind - unterschiedliche Kovarianzmatrizen besitzen (GORDON, 1981).

Das Ähnlichkeitsmaß ist $s_{rt} = \sum_{j=1}^{p} u_{rtj} / \sum_{j=1}^{p} v_{rtj}$, wobei für intervall- beziehungsweise verhältnisskalierte Variablen und äquidistante, ordinalskalierte Variablen die Gleichung

$$u_{rtj} = 1 - \left(\left| x_{rj} - x_{tj} \right| \right) / \text{Spannweite von } x_j$$

gilt; bei nominal- und nicht äquidistanten, ordinalskalierten Variablen nimmt u_{rtj} den Wert 1 an, wenn die Objekte r und t den gleichen Wert besitzen und den Wert 0 in allen anderen Fällen; bei binären Variablen nimmt u_{rtj} den Wert 1 an, wenn die Objekte r und t den Wert 1 besitzen und den Wert 0 in allen anderen Fällen. Während u_{rtj} ein Maß für die Ähnlichkeit von zwei Objekten ist, repräsentiert v_{rtj}, ob überhaupt ein Vergleich zwischen den Objekten möglich ist. Können r und t bei Variable j miteinander verglichen werden, so nimmt v_{rtj} den Wert 1 an, können sie nicht miteinander verglichen werden (zum Beispiel aufgrund fehlender Werte), wird v_{rtj} normalerweise gleich Null gesetzt. Für binäre Variablen ist v_{rtj} gleich Null, wenn bei Variable j sowohl bei r als auch bei t der Wert gleich Null ist. In allen anderen Fällen ist v_{rtj} = 1. Einen GOWER, 1971, vergleichbaren Ansatz haben KAUFMANN & ROUSSEEUW, 1990.

Im Bereich der mehrdimensionalen Skalierung werden in dieser Arbeit nur zwei Verfahren näher betrachtet und zwar die Hauptkoordinatenanalyse und die ordinale mehrdimensionale Skalierung.

Die Hauptkoordinatenanalyse, die bisweilen auch als klassische oder metrische mehrdimensionale Skalierung bezeichnet wird, geht zurück auf Arbeiten von TORGERSON, 1958, und GOWER, 1966. Wenn hier die Bezeichnung Hauptkoordinatenanalyse gewählt wird, so vor allem, um die, wie es GOWER, 1966, nennt, Dualität von Hauptkomponentenanlyse und Hauptkoordinatenanalyse, auszudrücken (siehe unten). Einige wichtige Gesichtspunkte der Hauptkoordinatenanalyse lassen sich wie folgt zusammenfassen:

1. Die gefundenen Achsen sind, wie in der Hauptkomponentenanlyse, rotierbar. Die Orientierung der gefundenen Konfiguration ist also nicht die einzig mögliche Lösung, und eine orthogonale Transformation oder auch eine Spiegelung, bei der Distanzen und Winkel erhalten bleiben, führen zu weiteren gültigen Darstellungen der (n x n) Proximitätsmatrix **D** (MORRISON, 1990).
2. Wie in der Hauptkomponentenanlyse werden die 'wichtigsten' Dimensionen durch die größten Eigenwerte repräsentiert und analog zur Hauptkomponentenanlyse läßt sich $G = \sum_{i=1}^{q} l_i / \sum_{i=1}^{n-1} l_i$ mit l_i als Eigenwert von Dimension i (i = 1 ... n) als Maß der Anpassungsgüte des reduzierten Modells in q Dimensionen ($q \leq n - 1$) berechnen (KRZANOWSKI, 1988a).

3. Solange $\mathbf{B} = \mathbf{X}\mathbf{X}'$ positiv semidefinit ist, ist G ein guter Anhaltspunkt für die Anpassungsgüte. **B** ist immer dann positiv semidefinit, wenn das Proximitätsmaß die metrische Ungleichung $d_{rt} \leq d_{ru} + d_{ut}$ erfüllt[11]. Sie wird von einer Reihe von Proximitätsmaßen, zum Beispiel der euklidischen Distanz, aber auch zum Beispiel dem Simple Matching-Ähnlichkeitsmaß erfüllt (GOWER, 1971).
4. Ist die metrische Ungleichung nicht erfüllt, wird **B** in der Regel nicht positiv semidefinit sein, und einer oder mehrere negative Eigenwerte vorliegen, das heißt, mindestens eine Dimension der Konfiguration ist imaginär, die Anpassungsgüte ist unbefriedigend und G wird überschätzt (SIBSON, 1979).
5. Eine weitere Möglichkeit die Anpassungsgüte zu beurteilen, ist ein Plot (der sogenannte Shepard Plot) der Proximitätswerte der Ausgangsmatrix gegen die euklidischen Distanzen der Objekte, der durch die Hauptkoordinatenanalyse in q Dimensionen ($q \leq n$) ermittelten Konfiguration. Eine gute Übereinstimmung wird durch einen linearen, durch den Ursprung gehenden Verlauf gekennzeichnet. Ein weiteres Hilfsmittel zur Überprüfung der Anpassungsgüte einer durch Dimensionserniedrigung erzielten Konfiguration ist die Überlagerung der dimensionserniedrigten Darstellung der Objekte durch einen Minimum Spanning Tree[12] (CHATFIELD & COLLINS, 1980).
6. Zwischen Hauptkoordinatenanalyse und Hauptkomponentenanalyse besteht eine Dualität, das heißt, die Hauptkoordinatenanalyse ergibt für die Objekte exakt dieselben Koordinaten aus **D** wie die Hauptkomponentenanlyse in Form der Hauptkomponentenwerte aus **X**, wenn die Proximitätsmatrix **D** die Matrix der euklidischen Distanzen der Objektmatrix **X** ist (GOWER, 1966). Die Hauptkoordinatenanalyse ist daher vor allem dann sinnvoll, wenn eine Proximitätsmatrix als Ausgangsmatrix vorliegt, oder die aus der Ausgangsmatrix **X** abgeleitete Proximitätsmatrix nicht durch die Bildung euklidischer Distanzen der Objekte von **X** gewonnen wird, sondern durch Anwendung eines anderen - der metrischen Ungleichung entsprechenden - Proximitätsmaßes, zum Beispiel bei Verwendung des allgemeinen Ähnlichkeitskoeffizienten aufgrund des Vorliegens von Variablen mit unterschiedlichen Skalenniveaus.

Während die Hauptkoordinatenanalyse bestrebt ist, die Unähnlichkeiten zwischen den Objekten der Ausgangsmatrix numerisch so exakt wie möglich abzubilden, wird in der ordinalen mehrdimensionalen Skalierung, die bisweilen auch als nicht-metrische mehrdimensionale

---

[11] Zur Verwendung der Bezeichnungen und Indices: d steht für ein Unähnlichkeitsmaß; die Indices r, u und t kennzeichnen drei Objekte.

[12] Minimum Spanning Trees werden gesondert in 2.5.1.2 angesprochen.

Skalierung bezeichnet wird, lediglich gefordert, daß die Rangfolge der Unähnlichkeiten der Ausgangsmatrix, der Rangfolge der Unähnlichkeiten, die durch die ordinale mehrdimensionale Skalierung erzielt wird, entspricht. Ein weiterer wichtiger Unterschied zwischen Hauptkoordinatenanalyse und ordinaler mehrdimensionaler Skalierung ist darüber hinaus, daß die Überprüfung der Anpassungsgüte ein integrierter Bestandteil der ordinalen mehrdimensionalen Skalierung ist; auf iterativem Weg wird in der ordinalen mehrdimensionalen Skalierung ein gewähltes Kriterium, und damit die Anpassungsgüte, optimiert. Neben den bereits im vorangegangenen Kapitel genannten Einsatzgebieten, spielt die ordinale mehrdimensionale Skalierung zusätzlich vor allem dort eine Rolle, wo die Daten in Form von Rängen vorliegen, oder zwar numerische Proximitäten vorliegen, diese aber mit viel Ungenauigkeit behaftet sind und/oder davon ausgegangen wird, daß letztlich auch die Rangfolge der Proximitäten ausreichend Informationen für die zu beantwortenden Fragestellungen beinhaltet.

Die ordinale mehrdimensionale Skalierung geht zurück auf die Veröffentlichungen von SHEPARD, 1962a und 1962b und KRUSKAL, 1964a und 1964b. Einen ausführlichen Überblick zur Methodik sowie Diskussionen liefern unter anderem SCHIFFMAN et al., 1981, SHEPARD et al., 1972, oder YOUNG, 1987.

Zur Methodik der einfachen, ordinalen mehrdimensionale Skalierung, einige einleitende Definitionen:

1. Als Dissimilaritäten δ_{rt} werden die Unähnlichkeiten der Ausgangsproximitätsmatrix bezeichnet;
2. als Distanzen d_{rt} werden die euklidischen Distanzen der Objekte in der durch die mehrdimensionale Skalierung erzielten Konfiguration in q Dimensionen bezeichnet;
3. als Disparität $\hat{d}_{rt}$ wird der Schätzwert bezeichnet, der durch die dem mehrdimensionalen Skalierungsmodell zugrunde gelegte Beziehung von δ_{rt} und d_{rt} geschätzt wird.

Bei der Durchführung einer ordinalen mehrdimensionalen Skalierung sind zu beachten:

1. Wahl der Anzahl der zu betrachtenden Dimensionen q; zu beachten ist, daß die Objektkoordinaten der ersten und zweiten Dimension einer Lösung in zwei Dimensionen nicht den Koordinaten der ersten und zweiten Dimension einer Lösung in drei Dimensionen entsprechen müssen; das heißt, die Anzahl der Dimensionen, in denen die Anpassungsgüte optimiert wird hat einen Einfluß auf die Koordinaten der zu betrachtenden Dimensionen. Vor allem q = 2 ist natürlich mit Hinblick auf die graphische Darstellung der Lösung vorteilhaft.
2. Festlegung der Ausgangskonfiguration in q Dimensionen; häufig wird als Ausgangskonfiguration das Ergebnis einer Hauptkoordinatenanalyse gewählt. Der Beginn mit mehreren (zufälligen) Ausgangskonfigurationen wird ebenfalls

empfohlen, um die Gefahr zu verringern, den Iterationsprozeß an einem lokalen, statt an dem globalen Minimum des Optimierungskriteriums zu beenden (GENSTAT COMMITTEE, 1993).

3. Wahl des Kriteriums, das optimiert werden soll; zum Beispiel

$$\text{stress} = \left[\sum_{r<t} \left(d_{rt} - \hat{d}_{rt} \right)^2 / \sum_{r<t} d_{rt}^2 \right]^{1/2}, \quad \text{logstress} = \sum_{r<t} \left(\log d_{rt} - \log \hat{d}_{rt} \right)^2,$$

oder $\text{non linear mapping stress} = \sum_{r<t} \left(d_{rt} - \delta_{rt} \right)^2 / \delta_{rt}$ (EVERITT, 1978, GENSTAT COMMITTEE, 1993).

4. Wiederholen der Schritte 1. bis 3. für mehrere Werte von q, in der Regel $1 \leq q \leq 5$. Beginn üblicherweise mit q_{max}. Möglich ist dann Verwendung dieser Konfiguration (in $q_{max} - 1$ Dimensionen) als Ausgangskonfiguration für die Analyse in der nächst niedrigen Dimensionszahl und entsprechend für alle weiteren Werte von q.

Einige zusätzliche Anmerkungen:

1. Die Festlegung der Anzahl der zu betrachtenden Dimensionen ist ähnlich problematisch wie in der Hauptkomponentenanalyse.
2. Je mehr Dimensionen betrachtet werden, desto geringer ist der Wert des stress-Kriteriums am Ende des Iterationsprozesses. Ab welchem Punkt jedoch ein ausreichend niedriger Wert erreicht ist, ist letztlich eine subjektive Entscheidung. Einige grobe Faustzahlen gibt KRUSKAL, 1964b, die jedoch nicht mehr als (ungenaue) Anhaltspunkte sein können[13].
3. Neben dem stress-Wert ist auch der Wert des quadrierten Korrelationskoeffizienten zwischen Distanzen und Disparitäten ein guter, oft sogar besserer, Anhaltspunkt für die Anpassungsgüte des Modells und die Festlegung der Anzahl der zu betrachtenden Dimensionen (SCHIFFMAN et al., 1981).
4. Da die Reproduktion der Ränge mehr Spielräume als die exakte, numerische Reproduktion läßt, ist mit der ordinalen mehrdimensionalen Skalierung häufig eine befriedigende Lösung in weniger Dimensionen zu finden als mit der Hauptkoordinatenanalyse. Allerdings sind sehr extreme Objekte in der Regel wenig stabil, das heißt sie können praktisch an einer beliebig fernen Stelle der Konfiguration plaziert werden (KRZANOWSKI, 1988a).

---

[13] Danach gilt zum Beispiel ein stress-Wert von 0,1 als befriedigend, von 0,05 als gut und von 0,025 als exzellent.

5. Vergleiche zwischen Hauptkoordinatenanalyse und ordinaler mehrdimensionaler Skalierung liefern zum Beispiel CHATFIELD & COLLINS, 1980; einen kurzen Überblick gibt GORDON, 1981. Verallgemeinernd läßt sich festhalten, daß in der Regel die ordinale mehrdimensionale Skalierung vergleichbare oder bessere und nur selten schlechtere Ergebnisse als die Hauptkoordinatenanalyse erbringt[14].

Als Kriterien für die Anzahl der zu betrachtenden Dimensionen nennt SHEPARD, 1972,

- die von KRUSKAL, 1964b, genannten Anhaltswerte des stress-Kriteriums;
- die Lage des 'Bruchs' im Diagramm der stress-Werte gegen die Anzahl der Dimensionen im Sinne der Argumentation des Scree-Diagramms (siehe Tabelle 1);
- die Interpretierbarkeit der Konfiguration;
- die Ähnlichkeitkeit von Konfigurationen aus Wiederholungen.

SCHIFFMAN et al., 1981, betonen daneben den Wert der quadrierten Korrelationen als gute Richtlinie, und RAMSAY, 1982, schlägt einen Signifikanztest vor. Darüber hinaus sind in diesem Zusammenhang die Arbeiten von KLAHR, 1969, LEVINE, 1978, und SPENCE, 1979, zu nennen. Als Hilfsmittel für die Entscheidung der zu betrachtenden Dimensionalität werden hier stress-Werte zufällig generierter Proximitätsmatrizen herangezogen. Liegen die stress-Werte einer aktuellen Untersuchung deutlich (30 - 50 %) unter den stress-Werten von Zufalls-Proximitätsmatrizen, so kann nach SPENCE, 1979, von nicht nur auf Zufallsvariabilität beruhenden Daten ausgegangen werden. Als Ergebnis der Untersuchungen von KLAHR, 1969 und LEVINE, 1978, ist festzuhalten:

1. je größer die Anzahl der Objekte ist, desto unwahrscheinlicher ist es, bei Zufalls-Proximitätsmatrizen geringe (das heißt unter 0.1) stress-Werte zu bekommen;
2. je größer die Anzahl der Objekte ist, desto ähnlicher werden sich die stress-Werte der Zufalls-Proximitätsmatrizen in einer zunehmenden Anzahl von Dimensionen (geprüft bis q = 5) und desto geringer wird die Abnahme des stress-Wertes bei Hinzunahme einer weiteren Dimension. Als Faustzahl werden 10 Objekte genannt. Sie sollten einer mehrdimensionalen Skalierung mindestens zur Verfügung stehen. Bei weniger als 10 Objekten ist die Gefahr groß, auch bei Daten ohne Struktur eine Struktur aufgrund eines niedrigen stress-Wertes zu vermuten.

Eine Approximation an den stress-Wert von Zufalls-Proximitätsmatrizen gibt SPENCE, 1979.

Abschließend soll kurz auf die Diskussion eingegangen werden, welche Ausgangskonfiguration bei einer ordinalen mehrdimensionalen Skalierung verwendet werden sollte. SPENCE, 1972, argumentiert für eine geplante (rationale) Startkonfiguration, vor allem mit dem Hinweis auf zu

---

[14] Besser und schlechter im Sinne der Anpassung der Konfiguration nach der Analyse an die wahre Konfiguration der Ausgangsdaten.

sparende Rechenzeit. Als mögliche rationale Startkonfiguration erwähnen SPENCE & YOUNG, 1978, zum Beispiel die Konfiguration, die durch eine Hauptkoordinatenanalyse erzielt wird. Die Gefahr an einem lokalen Minimum 'gefangen' zu werden schätzen sie bei dieser Strategie als relativ gering ein. Die rationale Ausgangskonfiguration wird vor allem als vorteilhaft gegenüber des Analysebeginns mit einer einzigen Zufalls-Ausgangskonfiguration angesehen. ARABIE, 1973, 1978a und 1978b, dagegen begründet die Vorteilhaftigkeit der Verwendung einer Zufallskonfiguration wie folgt:

1. es ist nicht klar, welche der möglichen rationalen Startkonfigurationen die beste ist im Hinblick auf die Vermeidung von Lösungen an lokalen Minima beziehungsweise dem Erzielen von Lösungen mit minimalen stress-Werten;
2. wenn demnach mehrere rationale Ausgangskonfigurationen verwendet werden sollen, sind größerer Rechenaufwand und größere Programmresourcen notwendig, als wenn eine beliebige Anzahl (ARABIE, 1978a empfiehlt 20) von Zufalls-Ausgangskonfigurationen erzeugt und analysiert wird, und die Lösung derjenigen Zufalls-Ausgangskonfiguration verwendet wird, die den geringsten stress-Wert erreicht.

Tabelle 2: Überblick über einige Proximitätsmaße

Skalenniveau der Variablen	Bezeichnung und Formel (Quellen: BACHER, 1994, EVERITT, 1980, GOWER & HAND, 1996, SCHUBÖ et al., 1991)
intervall- beziehungsweise verhältnisskaliert	Minkowski $d_{rt} = \left[\sum_{j=1}^{p} \left| x_{rj} - x_{tj} \right|^{R} \right]^{1/R}$ (wenn R = 1 enspricht dies der City-Block-Distanz, wenn R = 2 der euklidischen Distanz) Czekanowski $d_{rt} = 1 - \left[2\sum_{j=1}^{p} \min\left(x_{rj} - x_{tj}\right) / \sum_{j=1}^{p} \left(x_{rj} - x_{tj}\right) \right]$ Canberra $d_{rt} = \sum_{j=1}^{p} \left(\left| x_{rj} - x_{rt} \right| \right) / \left(x_{rj} - x_{rt} \right)$ Mahalanobis $d_{rt} = \sum_{j1=1}^{p} \sum_{j2=1}^{p} \left(x_{rj1} - x_{tj1} \right) v_{j1j2} \left(x_{rj2} - x_{tj2} \right)$ (j1 und j2 sind zwei von p Variablen und v_{j1j2} das Element der j1-ten Zeile und der j2-ten Spalte der Inversen der Kovarianzmatrix der p Variablen)
ordinalskaliert	bei äquidistanter Ordinalskala ist Verwendung der für intervall- beziehungsweise verhältnisskalierten Variablen entwickelten Proximitätsmaße möglich; wegen seiner sinnvollen ordinalen Interpretation ist vor allem das City-Block-Distanzmaß geeignet (siehe BACHER, 1994); bei nicht äquidistanter Ordinalskala werden für nominalskalierte Variablen entwickelten Proximitätsmaße verwendet.
nominalskaliert	Sneath Matching $d_{rt} = 1 - c_{rt}$ (c_{rt} ist gleich der Anzahl der Übereinstimmungen bei r und t, geteilt durch p)
binäre Variablen	Simple Matching $s_{rt} = (a + d) / (a + b + c + d)$ Jaccard $s_{rt} = a / (a + b + c)$ Roger & Tanimoto $s_{rt} = (a + d) / (a + d + 2(a + b))$ im Fall binärer Variablen gilt die folgende 2-Wege Tafel Objekt r Objekt t 1 0 1 a b 0 c d a+b+c+d = p, das heißt die Kontingenztafel zeigt auf, bei wieviel Variablen zwischen r und t Übereinstimmung (bei a und d) beziehungsweise nicht Übereinstimmung (bei b und c) besteht.

Bezeichnungen und Indices: d steht für ein Unähnlichkeits-, s für ein Ähnlichkeitsmaß; j steht für eine Variable, die Indices r und t kennzeichnen zwei Objekte; x_{rj} ist somit der Wert von Variable j bei Objekt r.

2.1.3 Korrespondenzanalyse

Die Korrespondenzanalyse ist eine weitere, überwiegend deskriptiv eingesetzte Methode zur graphischen Abbildung von Datenmatrizen, aufbauend auf der Eigenwertzerlegung (singular value decomposition) der Datenmatrix (GOOD, 1969). Die Korrespondenzanalyse ist vor allem von der französischen Statistik begründet und entwickelt worden (zum Beispiel BENZECRÍ, 1973). Ursprünglich stand die Analyse von Häufigkeitsdaten nominalskalierter Variablen in Form einer bivariaten Korrespondenzanalyse im Vordergrund. Entsprechende Kodierung ermöglicht aber auch die Analyse intervall-, beziehungsweise verhältnisskalierter Variablen, ordinalskalierter Variablen und gemischter Variablensätze. Einführende Darstellungen liefern zum Beispiel GREENACRE, 1981, HILL, 1974, oder JAMBU, 1991, zusammenfassende Gesamtdarstellungen GREENACRE, 1984 und 1993. Auf die enge Verbindung von Korrespondenzanalyse und die Analyse von Kontingenztafeln mit Hilfe log-linearer Modelle sei hingewiesen (siehe zum Beispiel VAN DER HEIJDEN & DE LEEUWS, 1985 oder VAN DER HEIJDEN et al., 1989).

Die bivariate Korrespondenzanalyse dient zur Analyse einer (k x p) Datenmatrix **Z**, einer Kontingenztafel mit i = 1 ... k Zeilen (k Ausprägungen der nominalskalierten Zeilenvariablen z_k) und j = 1 ... p Spalten (p Ausprägungen der nominalskalierten Spaltenvariablen z_j). Die Vektoren der Zeilen- und Spaltensummen von **Z** sind der (k x 1) Spaltenvektor **z** beziehungsweise der (1 x p) Zeilenvektor **s**. $\mathbf{D}_z$ ist die Diagonalmatrix der Zeilen-, $\mathbf{D}_s$ die Diagonalmatrix der Spaltensummen von **Z**.

Ziel der Korrespondenzanalyse ist die Darstellung der Zeilen- und/oder Spaltenprofile im - wenn sinnvoll und ohne großen Informationsverlust möglich - zweidimensionalen Raum beziehungsweise allgemein im q-dimensionalen Raum (q < min(k,p)). Zu den Koordinaten für die Darstellung der Zeilen- und Spaltenprofile gelangt man über die Eigenwertzerlegung der doppelt gewichteten Matrix **Z**, also durch die Eigenwertzerlegung von $\mathbf{Z}^{\mathbf{w}} = \mathbf{D}_z^{-1/2}\mathbf{Z}\mathbf{D}_s^{-1/2} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}'$, wobei **U** die (k x p) Matrix der linken singulären Vektoren, **V** die (p x p) Matrix der rechten singulären Vektoren und $\mathbf{D}_\alpha$ die p-dimensionale Diagonalmatrix der singulären Werte von $\mathbf{Z}^{\mathbf{w}}$ sind. Die Koordinaten für die Zeilenprofile errechnen sich dann als die Elemente der (k x p) Matrix $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$, die Koordinaten für die Spaltenprofile als die Elemente der (p x p) Matrix $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$, mit $\mathbf{D}_r$ als Diagonalmatrix der Zeilen- und $\mathbf{D}_c$ als Diagonalmatrix der Spaltensummen von $(1/\mathrm{N})\mathbf{Z}$, mit N als der Gesamthäufigkeit (GENSTAT COMMITTEE, 1993).

Eine Besonderheit der Korrespondenzanalyse als Folge der doppelten Wichtung von **Z** ist das Auftreten einer Lösung, die immer mit dem größten singulären Wert (= 1) korrespondiert und dessen Zeilen- und Spaltenkoordinaten ebenfalls gleich 1 sind. Der erste singuläre Wert und die mit ihm korrespondierenden singulären Vektoren werden daher in der Regel verworfen.

Die Koordinaten sind so skaliert, daß gelten $\mathbf{F}'\mathbf{D}_r\mathbf{F} = \mathbf{D}_\alpha^2$ und $\mathbf{G}'\mathbf{D}_c\mathbf{G} = \mathbf{D}_\alpha^2$. Andere

Skalierungen sind natürlich denkbar. Häufig verwendet wird eine Tranformation zu einer Standardisierung zur Einheitshauptachse durch $\mathbf{\Phi} = \mathbf{F}\mathbf{D}_{\alpha}^{-1}$ mit $\mathbf{\Phi}'\mathbf{D}_{r}\mathbf{\Phi} = \mathbf{I}$ und $\mathbf{\Gamma} = \mathbf{G}\mathbf{D}_{\alpha}^{-1}$ mit $\mathbf{\Gamma}'\mathbf{D}_{c}\mathbf{\Gamma} = \mathbf{I}$, wobei **I** eine (k x k) beziehungsweise (p x p) Einheitsmatrix ist. Diese Koordinaten werden dann auch als Standardkoordinaten bezeichnet.

In gewissen Fällen kann es informativ sein, die Zeilenprofile in Normal- und die Spaltenprofile in Standardkoordinaten darzustellen und umgekehrt. So führt zum Beispiel eine Darstellung der Zeilenprofile in Normal- und der Spaltenprofile in Standardkoordinaten zu einer Abbildung, in der der Zeilenprofilpunkt exakt am Zentroid der Spaltenprofilpunkte liegt, die das Zeilenprofil definieren. Eine Abbildung der Zeilenprofilpunkte in Standard- und der Spaltenprofilpunkte in Normalkoordinaten hingegen führt zu einer Abbildung, in der jeweilige Spaltenprofilpunkt am Zentroid der Zeilenprofilpunkte liegt, die der Kategorie des betrachteten Spaltenprofils zugerechnet werden können.

Zur Interpretation der Ergebnisse einer (bivariaten) Korrespondenzanalyse ist die graphische Abbildung der Profile der wichtigste Ausgangspunkt. Es ist zu beachten, daß die durch die Korrespondenzanalyse abgebildeten Chi-Quadrat Distanzen, die durch die euklidischen Distanzen in der dimensionserniedrigten Darstellung approximiert werden, nur innerhalb der Zeilenprofile, beziehungsweise nur innerhalb der Spaltenprofile als solche interpretiert werden dürfen. Die Distanz zwischen einem Zeilenprofilpunkt und einem Spaltenprofilpunkt ist dagegen nicht durch eine entsprechende Chi-Quadrat Distanz definiert. Neben der graphischen Abbildung der Zeilen- und Spaltenprofile sind folgende Kennwerte hervorzuheben:

1. Die singulären Werte beziehungsweise die Eigenwerte jeder Dimension (in der Sprache der Korrespondenzanalyse die Inertia jeder Dimension) sowie der Anteil der Inertia der betrachteten Dimensionen an der Gesamtinertia der Ausgangsmatrix.
2. Der absolute Beitrag eines Profilpunktes zur Definition der jeweiligen Dimension. Der absolute Beitrag gibt Auskunft darüber, in wie weit der jeweilige Profilpunkt an der Ausrichtung der jeweiligen Achse beteiligt ist und ist daher - vergleichbar mit den Koeffizienten (den Eigenvektoren) in der Hauptkomponentenanlyse - ein Anhaltspunkt für die Beschreibung und Interpretation der jeweils betrachteten Dimension.
3. Der relative Beitrag einer Dimension zur Inertia eines bestimmten Profilpunktes. Der relative Beitrag ist ein Maß für die Güte der Repräsentation eines Profils bei der gewählte Dimensionalität. Es ist durchaus denkbar, daß ein Profilpunkt zwar einen geringen absoluten Beitrag zur Ausrichtung der jeweiligen Dimension liefert, dennoch aber gut durch die gewählte Darstellung repräsentiert wird, das heißt einen hohen relativen Beitrag besitzt. Der relative Beitrag kann auch als der Winkel

zwischen den jeweiligen Achsen und einer den Ursprung und den Profilpunkt verbindenden Linie betrachtet werden.

4. Die Summe der relativen Beiträge, die auch als Qualität bezeichnet wird und deren Maximum 1 ist. Sie gibt Aufschluß über die Gesamtanpassung der Profile bei der gewählten Dimensionalität.

GREENACRE, 1993, schlägt darüber hinaus vor, für die Zeilen beziehungsweise Spalten Achsen zu berechnen und diese mit entsprechenden Markern zu versehen, die für die relativen Häufigkeiten stehen. GREENACRE, 1993, bezeichnet dieses Vorgehen als Kalibrierung. Eine orthogonale Projektion, zum Beispiel eines Zeilenprofilpunktes auf die so gebildete Achse einer Spaltenvariablen, ermöglicht das Abschätzen der relativen Häufigkeit der entsprechenden Zeilen-Spalten-Kombination. Voraussetzung ist natürlich eine gute Qualität der graphischen Repräsentation. In Zusammenhang mit den noch zu besprechenden Biplots (Kapitel 2.2) wird auf ähnliche Verfahrensweisen, im Bereich der multiplen Korrespondenzanalyse, näher eingegangen.

Die Kontingenztafel **Z** kann in die Indikatormatrix $\mathbf{Z}^{\mathbf{I}}$ umgewandelt werden, indem für jedes Objekt eine Zeile gebildet wird, und die Ausprägungen der Variablen in die Spalten geschrieben werden. Für Variable 1 (zum Beispiel die Spaltenvariable von **Z**) ergeben sich $j_1 = 1 \ldots p_1$ Spalten, für die Variable 2 (zum Beispiel die Zeilenvariable von **Z**) ergeben sich $j_2 = 1 \ldots p_2$ Spalten. Bei Zutreffen der jeweiligen Ausprägung einer Variablen bei Objekt i (i = 1 ... n) wird die Spalte, die dieser Ausprägung entspricht, mit einer 1, bei Nichtzutreffen mit einer 0 gekennzeichnet. Die so entstandene Indikatormatrix wird der Korrespondenzanalyse unterzogen und liefert in Bezug auf die relative Lage der Variablenkoordinaten dieselbe Darstellung wie die Korrespondenzanalyse der Häufigkeitsmatrix. Die Werte der Gesamtinertia der Achsen sind bei Analyse von $\mathbf{Z}^{\mathbf{I}}$ jedoch in der Regel größer als bei Analyse von **Z**, die Achsen sind im Vergleich gestaucht oder gestreckt. Für die Zeilen (Objekte) können wie für die Spaltenvariablen Koordinaten errechnet werden, wobei die Objekte mit identischen Werten auf einem Punkt zusammenfallen.

An Stelle von lediglich zwei Variablen kann eine Vielzahl von Variablen betrachtet werden und es wird folglich von einer multivariaten Indikatormatrix und entsprechend von einer multivariaten oder auch häufig von einer multiplen Korrespondenzanalyse gesprochen. Wie in der bivariaten Korrespondenzanalyse spielen die graphische Darstellung der Variablen und der Objekte, sowie Aussagen zu absolutem und relativem Beitrag, zur Qualität und zur Inertia der einzelnen Dimensionen und Profilpunkte, eine wichtige Rolle. Die relativen Werte der Inertia der ersten (zwei) Dimensionen sind in der Regel gering. Als Hauptursache führt GREENACRE, 1991, die künstliche Schaffung zusätzlicher Dimensionen durch die oben beschriebene Kodierung an.

Die Bildung einer Indikatormatrix ermöglicht die Verrechnung von Variablen mit beliebigen Skalenniveaus und gemischter Variablensätze. Notwendig ist allerdings die Diskretisierung nicht nominalskalierter Variablen, das heißt ordinal-, intervall- und verhältnisskalierte Variablen müssen so kodiert werden, daß entsprechende Kategorien oder Klassen gebildet werden. Kodierungen

haben in der Regel einen Informationsverlust zur Folge und unterliegen subjektiven Entscheidungen. Es kann daher angebracht sein zu überprüfen, ob und wie stark unterschiedliche Klassenbildungen beziehungsweise Kodierungsverfahren die Analyseergebnisse beeinflussen.

Alternativen zur multiplen Korrespondenzanalyse der Indikatormatrix sind die multiple Korrespondenzanalyse der Burt-Matrix beziehungsweise die gemeinsame Korrespondenzanalyse (joint correspondence analysis). Die Burt-Matrix (BURT, 1950) $\mathbf{Z^B}$ berechnet sich als $\mathbf{Z^B} = \mathbf{Z^{I'} Z^I}$. Bei Verwendung von Standardkoordinaten ergibt die Analyse von $\mathbf{Z^B}$ Spaltenkoordinaten, die den (Standard-) Spaltenkoordinaten der Analyse von $\mathbf{Z^I}$ entsprechen. Unterschiede bei den Normalkoordinaten sind bedingt durch Unterschiede bei den Inertias (α) der Dimensionen, die wie folgt in Beziehung stehen: $\alpha^2_{\mathbf{Z^B}} = (\alpha^2_{\mathbf{Z^I}})^2$. Informationen zu den Objekten gehen bei Verwendung der Burt-Matrix natürlich verloren. Wo diese von besonderem Interesse sind, wie in der vorliegenden Arbeit, ist die Korrespondenzanalyse der Burt-Matrix daher nur eingeschränkt sinnvoll.

Die gemeinsame Korrespondenzanalyse (joint correspondence analysis) kann als Analyse der Elemente außerhalb der Diagonalen der Burt-Matrix verstanden werden (GREENACRE, 1988, 1991 & 1993, GOWER & HAND, 1996). Ihre Lösung erfolgt auf iterativem Weg. Durch die gemeinsame Korrespondenzanalyse wird eine gegenüber der multiplen Korrespondenzanalyse der Indikatormatrix verbesserte graphische Repräsentation der Beziehungen der Variablen untereinander erzielt. Zur Interpretation wird die Darstellung einer Variablen in Standardkoordinaten und die Darstellung der übrigen Variablen in Normalkoordinaten empfohlen. Die Standardkoordinaten der Kategorien der ausgewählten Variablen dienen dann als Referenzpunkte für die Interpretation der Beziehungen der übrigen Variablen zu der in Standardkoordinaten abgebildeten Variablen. Steht der iterative Algorithmus zur Durchführung einer gemeinsamen Korrespondenzanalyse nicht zur Verfügung, kann durch ein Reskalierungsverfahren das Ergebnis einer multiplen Korrespondenzanalyse der Burt-Matrix den Ergebnissen einer gemeinsamen Korrespondenzanalyse angenähert werden. Die Objekte betreffend gilt für die gemeinsame Korrespondenzanalyse dieselbe Einschränkung wie für die Korrespondenzanalyse der Burt-Matrix (siehe oben).

Häufig treten in gartenbaulichen Untersuchungen ordinalskalierte Variablen zum Beispiel in Form von Boniturwerten auf. Diese ordinalskalierten Variablen besitzen eine obere und eine untere Grenze und werden daher auch als bipolare Daten bezeichnet. Die Korrespondenzanalyse bipolarer Daten erfolgt durch Analyse der 'verdoppelten' Matrix (GREENACRE, 1984). Eine Matrix ordinalskalierter Variablen wird 'verdoppelt', indem für jede Variable eine Plus- und eine Minusspalte gebildet wird. Sind p ordinalskalierte Variablen gegeben (j = 1 ... p), und wird t_j als die obere Grenze der Bonitursкala von Variable j, und z_{ij} als der Boniturwert von Objekt i bei Variable j definiert

(i = 1 ... n), so errechnet sich die Plusspalte j+ als z_{ij} und die Minusspalte j- als $t_j - z_{ij}$. Die so 'verdoppelte' Matrix wird der Korrespondenzanalyse unterzogen. Im Korrespondenzanalyseplot fallen Ursprung und Zentroid der Koordinatenmatrix zusammen. Eine gedachte Linie vom Plus- zum Minuspol jeder Variablen führt durch den Ursprung. Die Distanz vom Ursprung zu den jeweiligen Koordinaten eines Plus- oder eines Minuspols (d_{j+} oder d_{j-}) ist in der vollen Dimensionalität gleich dem Variationskoeffizient der Plus- oder Minusspalte. Wird eine reduzierte Dimensionalität betrachtet, bieten d_{j+} und d_{j-} natürlich nur Approximationen an die Variationskoeffizienten.

Zwei weitere Kennwerte in der Interpretation der Korrespondenzanalyse bipolarer Daten sind die Polarisation des Mittels und die Polarisation der Objekte. Wenn $\tilde{k}_j = \bar{z}_j / t_j$, mit $\bar{z}_j$ als dem mittleren Boniturwert von Variable j definiert wird und andererseits $1 - \tilde{k}_j = (t_j - \bar{z}_j) / t_j$ ist, so ist das Produkt von $\tilde{k}_j$ und $(1 - \tilde{k}_j)$, also $\tilde{k}_j (1 - \tilde{k}_j)$ umgekehrt proportional zur sogenannten Polarisation des Mittels (pol_{mj}). Die geringste Polarisation des Mittels ergibt sich, wenn $\tilde{k}_j = 0.5$ (und damit $\tilde{k}_j (1 - \tilde{k}_j) = 0.25 = \text{Maximum}$) ist. pol_{mj} wird definiert als $pol_{mj} = 1 / \left(\tilde{k}_j (1 - \tilde{k}_j) \right)$ und ist immer ≥ 4. Je größer die Polarisation des Mittels ist, desto mehr Bedeutung kommt einem der beiden Extremwerte der Bonitурskala zu.

Eine hohe Polarisation der Objekte hingegen deutet auf die Lage der Objekte nahe den Pole, das heißt auf extreme Boniturwerte hin, während eine geringe Polarisation der Objekte auf dem Mittel nahe liegende Bonituren hinweist. Wenn $k_{ij} = z_{ij} / t_j$ mit z_{ij} als Wert von Objekt i (i = 1 ... n) bei Variable j (1 ... p) ist, so drückt ein hoher k_{ij}-Wert die Nähe des Objekts i zum Pluspol von Variable j und Ferne zum Minuspol derselben Variablen aus. Die Polarisation der Objekte berrechnet sich dann durch $pol_{ob} = n / \left(\sum_{i=1}^{n} k_{ij} (1 - k_{ij}) \right)$. Wie die Polarisation des Mittels hat sie Minimum von 4.

Die durch den Ursprung gehende (gedachte) Linie vom Punkt des positiven, zum Punkt des negativen Pols jeder Variablen - die der Summe von d_{j+} und d_{j-} entspricht - kann wie folgt interpretiert werden.

1. Ist das Verhältnis der größeren zu der kleineren Distanz d_{j+} und d_{j-} für die Variablen (annähernd) gleich, so ist auch die Polarisation des Mittels annähernd gleich, so daß die Gesamtlänge der Linie proportional zu S_j / t_j, das heißt proportional zur mit der oberen Grenze der Bonitурskala gewichteten Standardabweichung (S_j) dieser Variablen, ist.
2. Ist die Summe d_{j+} und d_{j-} (annähernd) gleich, aber das Verhältnis von größerer zu kleiner Distanz unterschiedlich, liegt also unterschiedliche Polarisation des Mittels vor, so heißt das, daß die am geringsten polarisierte Variable die größte Standardabweichung besitzt.

3. Sind sowohl die Gesamtlänge als auch das Verhältnis der Distanzen unterschiedlich, so ist die Gesamtlänge das Ergebnis einer Wechselwirkung von Standardabweichung und Polarisation des Mittels und nimmt mit steigender Standardabweichung und steigender Polarisation des Mittels zu.

Der Kosinus des Winkels zwischen zwei Linien zweier Spalten approximiert die Korrelation zwischen diesen Spalten.

Die Koordinaten der Zeilen (Objekte) der 'verdoppelten' Matrix sind äquivalent der Hauptkomponentenwerte dieser Objekte, bei Durchführung einer Hauptkomponentenanlyse der einfachen, nicht 'verdoppelten' Matrix, wenn die Variablen derart tranformiert werden, daß gilt $z^{*}_{ij} = \sqrt{fac_j}\, z_{ij}$ mit dem Faktor fac_j als $fac_j = (t_j \sum_{j=1}^{p} t_j) / \left[\bar{z}_j (t_j - \bar{z}_j)\right]$. Diese Transformation führt im Vergleich zur häufig in der Hauptkomponentenanalyse durchgeführten Standardisierung (siehe 2.1.1) zu einer stärkeren Hervorhebung stark polarisierter Objekte. Je größer die Polarisation eines Objekts ist, desto stärker geht sie in die Berechnung der Distanzen zweier Objekte ein. *Ein Genstat Code für die Korrespondenzanalyse bipolarer Daten ist im Anhang Teil III zu finden.*

Ist durch die Korrespondenzanalyse einer Matrix eine Abbildung der Häufigkeits-, Indikator-oder 'verdoppelten' Matrix erstellt, kann es informativ sein, in die vorhandene Darstellung zusätzliche Punkte, sei es Zeilen- oder Spaltenprofile, miteinzubeziehen. Solche zusätzlichen Punkte können, zum Beispiel in der gruppenweisen Analyse, die Ergebnisse anderer Objekte, oder auch externe Variablen sein. Die Koordinaten der zusätzlichen Punkte erhält man durch Anwendung geeigneter Transitionsformeln. Auf graphischem Weg ist diese Interpolation ebenfalls möglich (siehe Kapitel 2.2).

Abschließend einige Anmerkungen zur Beurteilung der Stabilität von Korrespondenzanalyse-Lösungen[15]. GREENACRE, 1984, unterscheidet zwischen interner und externer Stabilität. Der Begriff interne Stabilität bezieht sich auf die Ausgangsmatrix an sich, das heißt die interne Stabilität beurteilt, wie stark die Repräsentation der Matrix in der Korrespondenzanalyse von einzelnen Objekten beeinflußt wird. Sind Ausreißer oder Objekte mit sehr großer Leverage vorhanden, so kann die Entfernung dieser Objekte aus der Ausgangsmatrix die Repräsentation der Matrix erheblich verändern, die Lösung ist also intern instabil. Auch die Bedeutung einzelner Variablen wird als Merkmal interner Stabilität verstanden.[16]

---

[15] Grundsätzlich sind diese Gedanken natürlich auch auf die Lösungen anderer dimensionserniedrigender Analysen übertragbar.

[16] Darüberhinaus spielt die Selektion der Variablen mit der größten Bedeutung, ähnlich wie in der multiplen Regressions- oder linearen Diskriminanzanalyse, eine Rolle. KRZANOWSKI, 1993, stellt verschiedene Verfahren zur Variablenselektion in der Korrespondenzanalyse vor, auf die in dieser Arbeit jedoch nicht eingegangen werden soll.

Der Begriff externe Stabilität bezieht sich auf die Beziehung der Ausgangsmatrix zu der multivariaten Grundgesamtheit, aus der sie als Stichprobe ermittelt wurde. Werden weitere Stichproben gezogen, die zu stark abweichenden Lösungen führen, so ist die Lösung als extern instabil anzusehen.

Zur Beurteilung der internen Stabilität empfiehlt GREENACRE, 1984, Jackknifing, zur Beurteilung der externen Stabilität, Bootstrapping. Beide Verfahren werden in Kapitel 5 kurz angesprochen. *Eine interne Stabilitätsbeurteilung wird in Kapitel 3 eingesetzt und liegt als Genstat Code im Anhang Teil III vor.*

2.1.4 Faktoranalyse

Bei der Faktoranalyse handelt es sich wie bei der Hauptkomponentenanlyse um eine variablenorientierte R-Technik für intervall- beziehungsweise verhältnisskalierte und ordinalskalierte Variablen, die zu einer Dimensionserniedrigung genutzt werden kann. Sie wird ausführlich zum Beispiel von HARMAN, 1976, dargestellt. BARTHOLOMEW 1984 und 1985, unternimmt den Versuch, ein allgemeines Faktoranalysemodell zu definieren, das bei Variablen aller Skalenarten und bei gemischten Variablensätzen zu entsprechenden Lösungen führt. Ob überhaupt, und inwieweit, sich diese Ansätze gegenüber der 'klassischen' Faktoranalyse durchsetzen werden, ist noch unklar (siehe die Diskussion zu BARTHOLOMEW, 1985, zum Beispiel McDONALD, 1985). Obwohl eine gewisse Ähnlichkeit zur Hauptkomponentenanlyse vorhanden ist, gibt es auch wichtige Unterschiede. Im Gegensatz zur Hauptkomponentenanlyse liegt der Faktoranalyse ein gedankliches Konzept zugrunde und zwar das der latenten Variablen (beziehungsweise Faktoren). Dieses Konzept kann wie folgt erläutert werden. Die Korrelation $r_{x_i x_i^*}$ zwischen zwei Variablen x_i und x_i^* kann das Resultat ihrer gemeinsamen starken Korrelation mit einer weiteren Variablen y_j sein. Wenn dies zutrifft, ist die partielle Korrelation $r_{x_i x_i^* \bullet y_j}$ sehr gering, das heißt die Residualkorrelation zwischen x_i und x_i^* ist gering, nach Berücksichtigung des linearen Effekts von y_j auf sowohl x_i als auch x_i^*. Im Konzept der Faktoranalyse wird nun davon ausgegangen, daß es für die beobachteten Variablen, die in der Faktoranalyse auch als manifeste Variablen bezeichnet werden, eine (sehr viel geringere) Anzahl solcher latenter Variablen gibt, die zu eben diesem Effekt der Reduktion der partiellen Korrelation führen. Da diese Variablen aber nicht meß- oder beobachtbar sind, werden sie als latente Variablen bezeichnet. Sie können zwar nicht gemessen, wohl aber mit Hilfe des Faktoranalysemodells geschätzt werden. Daneben gibt eine Vielzahl weiterer, vor allem methodischer, Unterscheidungen, auf die hier nicht eingegangen wird. Eine Zusammenfassung wichtiger Gemeinsamkeiten und Unterschiede von Hauptkomponenten- und Faktoranalyse gibt Tabelle 3.

Kontrovers wird nach wie vor über die Vorzüglichkeit der einen gegenüber der anderen Methode diskutiert. VELICER & JACKSON, 1990a und 1990b geben einen Überblick über diese Diskussion. Als Hauptpunkte lassen sich festhalten:

1. Häufig sind die Lösungen sowohl im Bereich der Ladungen beziehungsweise Koeffizienten als auch im Bereich der geschätzten Faktorwerte beziehungsweise Hauptkomponentenwerte hoch korreliert.
2. Wo Ergebnisse stark unterschiedlich sind, liegt in der Regel eine Überextraktion von Faktoren beziehungsweise Hauptkomponenten und/oder ein schlecht definiertes Modell vor.
3. Impropere Lösungen (negative Schätzer für Elemente der Residuen (in der Sprache der Faktoranalyse der spezifischen Varianzen oder spezifischen

Faktoren)) stellen in der Faktoranalyse ein Problem dar. Sie müssen aber nicht nur als negativ angesehen werden, da sie als Diagnoseinstrument dienen können, um die Angemessenheit des Faktoranalysemodells zu überprüfen. Ausschalten improperer Lösungen durch einfache Manipulationen wie Begrenzung der Fehlerterme auf größer 0 nutzen diese Möglichkeit natürlich nicht und sind daher abzulehnen.

4. Die Lösung der Hauptkomponentenanlyse ist einfacher und schneller als die der Faktoranalyse. Zwar nimmt die Fähigkeit der Rechner zu, gleichzeitig steigt aber auch die Komplexität und Dimensionalität der zu verarbeitenden Daten. Der Geschwindigkeitsunterschied wird daher im wesentlichen erhalten bleiben.
5. Die Faktor-Unbestimmtheit - nicht zu verwechseln mit der Rotations-Unbestimmtheit, die ja auch für die Hauptkomponentenanlyse gilt - stellt ein besonderes, nach wie vor ungelöstes Problem in der Faktoranalyse dar. Sie ist die Konsequenz der Tatsache, daß im Faktoranalysemodell mehr Parameter geschätzt werden als Ausgangsvariablen, und damit Freiheitsgrade, vorhanden sind; aus p Variablen sind Parameter für q gemeinsame Faktoren und p spezifische Varianzen zu schätzen. Faktor-Bestimmtheit ist nur gegeben, wenn $q_{max} = 1/2\left((p-q)^2 - (p+q)\right) \geq 0$ gilt. Werden mehr als q_{max} Faktoren extrahiert, hat dies eine unbestimmte, und schon vor Rotation, uneindeutige Lösung zur Folge.
6. Eine Trennung in rein deskriptive Hauptkomponentenanalyse und modellbegründetete, schließende Faktoranalyse, ist irreführend, da auch die Hauptkomponentenanlyse konfirmatorische Aspekte beinhaltet, wenn bestimmte Modellannahmen zutreffen.

Die Schlußfolgerung von VELICER & JACKSON, 1990a, ist, daß in vielen Fällen die Hauptkomponentenanlyse der Faktoranalyse vorzuziehen ist. Zu ähnlichen, bisweilen weit radikaleren Schlußfolgerungen, kommen auch HILLS, 1977, SCHÖNEMANN, 1990 und STEIGER, 1990. Die Anhänger der Faktoranalyse finden zum Beispiel in McARDLE, 1990, oder MULAIK, 1990, ihre Fürsprecher.

In dieser Arbeit wird auf die Anwendung der Faktoranalyse oder verwandter Methoden, die auf der Vorstellung von latenten Variablen beruhen verzichtet[17]. Das Zutreffen insbesondere der konzeptionellen Grundlagen der Faktoranalyse wird unter Berücksichtigung der noch zu

---

[17] Neben der Faktoranalyse spielt auch die latente Strukturanalyse nominalskaliereter Variablen (BARTHOLOMEW, 1980, LAZARSFELD & HENRY, 1968) und die Analyse linearere Strukturgleichungsmodelle (JÖRESKOG & SÖRENBOM, 1993, PFEIFFER & SCHMIDT, 1987) in der Analyse von Modellen mit latenten Variablen eine Rolle.

besprechenden Daten bezweifelt; die statistischen Modellannahmen der Faktoranalyse werden durch die vorliegenden Daten nicht gedeckt; angesichts der darüber hinaus nicht zu übersehenden theoretischen Probleme der Faktoranalyse (impropere Lösungen, Faktor-Unbestimmtheit, Schätzung der Faktorwerte) ist ein Rückgriff auf diese Methodik bei der expolorativen Zielsetzung dieser Arbeit nicht erforderlich.

Tabelle 3: Gemeinsamkeiten und Unterschiede von Faktoranalyse und Hauptkomponentenanalyse

	Faktoranalyse	**Hauptkomponentenanlyse**
Gemeinsamkeiten		
Technik	R-Technik	
Orientierung	variablenorientiert	
Rotation	Rotation der Ergebnisse zulässig	
Variablenarten	in erster Linie für intervall- und verhältnisskalierte Variablen	
Interpretation	Interpretation der Ladungen beziehungsweise Koeffizienten´wichtig, aber nicht unproblematisch	
Unterschiede		
Konzept	Konzept latenter Variablen	kein zugrundeliegendes gedankliches Konzept
Zielrichtung	Erklärung der Kovarianzstruktur	Beschreibung der Varianzstruktur
statistische Modellannahmen	Vielzahl von Annahmen für Lösung der MLFA[18] notwendig	Lösung ohne statistische Modellannahmen möglich
Skalierung der Ausgangsvariablen	Lösung unverändert bis auf konstanten Faktor	Skalierung beeinflußt Lösung
Werte der Objekte	Berechnung der geschätzten Faktorwerte nach verschiedenen Verfahren und nicht eindeutig möglich	Berechnung der Hauptkomponentenwerte unproblematisch
Anzahl der Faktoren	Koeffizienten (Ladungen) verändern sich mit Anzahl betrachteter Faktoren	ohne Einfluß auf Hauptkomponenten und Koeffizienten

[18] Maximum Likelihood Factor Analysis; zu Einzelheiten der Berechnung der latenten Variablen, spezifischen Varianzen und der Faktorwerte, sowie der Modellannahmen siehe zum Beispiel KRZANOWSKI, 1988a.

2.2 Biplots

Biplots sind graphische Darstellungen von Datenmatrizen, die gleichzeitig Objekte und Variablen in einer Graphik abbilden (daher auch 'Bi'plots). Biplots stellen demnach nicht eine eigene Analysemethode dar, sondern bieten die Möglichkeit der Visualisierung der Zeilen und Spalten einer Datenmatrix, aufbauend auf verschiedenen dimensionserniedrigenden Verfahren (zum Beispiel der Hauptkomponentenanalyse, der mehrdimensionalen Skalierung und der Korrespondenzanalyse).

Die in dieser Arbeit gewählte Erläuterung und Darstellung der Biplots basiert auf GOWER & HAND, 1996. Die Visualisierung der Datenmatrix in Biplots dieses Typus ermöglicht sowohl die graphische Interpolation neuer (das heißt nicht an der Konstruktion des Biplots beteiligter) Objekte, als auch die graphische Prediktion der Variablenwerte der vorhandenen Objekte. Sind also die Variablenwerte eines neuen Objekts bekannt, so läßt sich die Position des Objekts im Biplot abschätzen (Interpolation); ist auf der anderen Seite die Lage eines Objekts im Biplot bekannt, so lassen sich die Werte der Variablen für dieses Objekt vorhersagen (Prediktion).

Die Konstruktion eines Biplots erfolgt in der Art, daß zunächst durch ein geeignetes Verfahren der Dimensionserniedrigung die Koordinaten der Objekte im dimensionserniedrigten, vorzugsweise zweidimensionalen, Raum gefunden werden, und dann entsprechend der Zielrichtung (Interpolation oder Prediktion) in das neue Achsensystem, das als Referenzsystem dient, die Biplotachsen als (nicht orthogonale) Achsen der Variablen eingezeichnet werden. Die graphische, deskriptive Interpretation der Daten steht dann im Vordergund der Arbeit mit Biplots.

2.2.1 Hauptkomponentenanalyse-Biplots

Die Hauptkomponentenanalyse der (n x p) Datenmatrix **X** führt zur (p x p) Matrix der Eigenvektoren **A** und zur (n x p) Matrix der Hauptkomponentenwerte **Y**. Die Hauptkomponentenwerte liefern die Koordinaten der Objekte im q-dimensionalen Unterraum L des p-dimensionalen Ausgangsraumes R als orthogonale, also unkorrelierte, Projektionen in der Art, daß die quadrierten Abweichungen der Distanzen der Objekte in der q-dimensionalen Projektion von den Distanzen der Objekte im p-dimensionalen Raum, minimiert werden, das heißt es gilt: $\sum_{r<t}^{n}\left(d_{rt}^{p}\right)^2 = \sum_{r<t}^{n}\left(d_{rt}^{q}\right)^2 + n\sum_{i=1}^{n} r_i^2$,

wobei die Indizes r und t für zwei von n (i = 1 ... n) Objekten stehen, d_{rt}^{p}, die aus der Datenmatrix **X** abgeleitete euklidische Distanz der Objekte r und t ist, und d_{rt}^{q} die im q-dimensionalen Unterraum definierte euklidische Distanz zwischen r und t darstellt. r_i schließlich ist die Abweichung zwischen d_{rt}^{p} und d_{rt}^{q}, um deren Minimierung es letztlich geht.

Zusätzlich zu den Positionen der Objekte sind nun in Biplots die Variablenachsen, die auch als Biplotachsen bezeichnet werden, zu ermitteln. Diese Achsen weisen die folgenden Merkmale auf:

1. Jede Variablenachse ist mit Markern versehen, die die Werte der Variablen in den Orginaleinheiten der Variablen wiedergeben;
2. die Positionen der Marker sind so gewählt, daß je nach Zielsetzung der Analyse, eine Interpolation oder eine Prediktion möglich ist;
3. die Länge der Biplotachsen vom kleinsten zum größten Marker ist ein Maß für die Güte der Repräsentation der Variablen im Biplot. Je länger die Biplotachse im Interpolations-Biplot, desto besser ist die Repräsentation der betreffenden Variablen;
4. der Kosinus des Winkels zwischen zwei Biplotachsen approximiert die Korrelation zwischen den Variablen;
5. die Richtung der Biplotachsen ist ein Indiz für die Korrelation der Variablen mit den Hauptkomponenten.

Bei der Erstellung eines Biplots beziehungsweise von auf Skalierungsverfahren wie der Hauptkomponentenanalyse, mehrdimensionalen Skalierung und Korrespondenzanalyse beruhenden graphischen Abbildungen, die der Visualisierung von Distanzen dienen, ist zu beachten, daß auf den Achsen der Hauptkomponenten gleiche Maßstäbe verwendet werden, da sich nur dann eine realistische Interpretation der Objektdistanzen ergibt.

2.2.1.1 Berechnung der Biplotachsen und Marker

Ausgehend von einem geeigneten Wert für jede Ausgangsvariable, zum Beispiel die dem Mittelwert einer Variablen am nächsten gelegene ganze Zahl[19], kann $\bar{x}_j + \mathrm{val} = i$ berechnet werden (im ganzzahligen Beispiel gilt dann $0 \le \mathrm{val} \le 1$)[20].

Die Koordinaten für den Marker von *i* ergeben sich dann für die Interpolationsmarker als $\mathbf{m}_{\mathbf{int}}^{ij} = \mathrm{val}(\mathbf{e_j A_q})$ und für die Prediktionsmarker als $\mathbf{m}_{\mathbf{pred}}^{ij} = \mathrm{val}\left(\mathbf{e_j A_q} \,/\, \mathbf{e_j A_q A_q' e_j'}\right)$ wobei $\mathbf{m}^{ij}$ der (1 x q) Zeilenvektor der Koordinaten von *i* in q Dimensionen bei Variable j, $\mathbf{A_q}$ die (p x q) Matrix der Eigenvektoren mit q Dimensionen und $\mathbf{e_j}$ eine (1 x p) Matrix mit einer 1 an der Stelle der Variablen j (j = 1 ... p) und ansonsten nur Nullen sind. Die Werte links und rechts vom Ausgangswert *i* sind durch Multiplikation mit einer, dem gewählten Markerabstand entsprechenden Konstanten bis zum Einschluß der kleinsten und größten Variablenwerte zu finden, also durch $\mathbf{m}_{\mathbf{int}}^{\nu j} = \nu \mathbf{m}_{\mathbf{int}}^{\nu j}$ beziehungsweise $\mathbf{m}_{\mathbf{pred}}^{\nu j} = \nu \mathbf{m}_{\mathbf{pred}}^{\nu j}$, wobei ν die Werte $\pm 1, \pm 2$,... und so weiter annimmt, und der Index ν für den Marker beim entsprechenden Multiplikator steht.

---

[19] Je nach Spannweite und Variablenart ergeben sich hier entsprechende Werte zum Beispiel in 10er oder 100er Schritten.

[20] $\bar{x}_j$ Mittelwert von Variable j (j = 1 ... p)

2.2.1.2 Interpolation und Prediktion

Die rechnerische Interpolation und Prediktion kann durch entsprechende Formeln erfolgen (siehe zum Beispiel JACKSON, 1991). Möglich ist nun im Biplot eine graphische Interpolation beziehungsweise Prediktion. Die Interpolation erfolgt durch die sogenannte Vektorsummenmethode. Dabei sind die Variablenwerte auf den Biplotachsen des zu interpolierenden Objekts miteinander zu verbinden und der Zentroid des so gefundenen Polygons festzulegen. Die Entfernung vom Ursprung - im Hauptkomponentenanalyse-Biplot also vom gemeinsamen Schnittpunkt aller Biplotachsen - zum Zentroid dieses Polygons, ergibt, multipliziert mit der Anzahl der betrachteten Variablen in der, durch die Lage dieses Zentroids bestimmten Richtung, die interpolierte Position des neuen Objekts. Die Prediktion der Variablenwerte erfolgt durch orthogonale Projektion vom Objektpunkt auf die jeweiligen Variablenachsen. Es ist zu beachten, daß Interpolations- und Prediktionsmarker unterschiedliche Positionen auf den Biplotachsen einnehmen, und daher immer nur der für den jeweiligen Zweck bestimmte Biplot verwendet werden darf.

2.2.1.3 Güte der Variablenrepräsentation

Visuell läßt sich die Güte der Variablenrepräsentation bereits durch die Länge der Biplotachsen der einzelnen Variablen beurteilen. Aufbauend auf den Eigenvektorwerten der einzelnen Variablen, lassen sich auch sogenannte CUSUM Diagramme erstellen (ARNOLD & COLLINS, 1993). Es gilt: $\sqrt{l_{j^*}} = \sum_{j=1}^{p} \sum_{j^*=1}^{q} a_{jj^*}^2 \sqrt{l_{j^*}}$, wobei l_{j^*} der Eigenwert der j^*-ten Hauptkomponente ($j^* = 1 \ldots q$) und a_{jj^*} die Elemente der (p x q) Matrix der Eigenvektoren $\mathbf{A_q}$ sind. Der Beitrag con_{11} der ersten Variablen zum Eigenwert der ersten Hauptkomponente errechnet sich dann zum Beispiel nach $con_{11} = a_{11}^2 \sqrt{l_1}$. Die Aufaddierung der Beiträge der einzelnen Variablen ergibt die Eigenwerte der einzelnen Hauptkomponenten. Die Abbildung der Beiträge in Form kumulativer Balkendiagramme mit den Beiträgen auf der Ordinate und den Hauptkomponenten auf der Abszisse, ermöglicht einen gleichzeitigen Einblick in die Bedeutung der Dimensionen und die Beiträge der Variablen.

Es ist festzuhalten, daß die Hauptkomponentenanalyse-Biplots in erster Linie auf die der Hauptkomponentenanalyse der Kovarianzmatrix aufbauen. Natürlich ist auch die Analyse der Korrelationsmatrix möglich, jedoch verliert der Hauptkomponentenanalyse-Biplot dann das wünschenswerte Merkmal der direkten Ablesbarkeit der Variablen-Orginalwerte und verwendet an dessen Stelle die standardisierten Werte. *Der Anhang enthält in Teil III Genstat Codes zur Erstellung von Hauptkomponentenanalyse-Biplots mit Interpolations- und Prediktionsmarkern, inklusive der Möglichkeit der interaktiven Prediktion bei Verwendung standardisierter Daten, sowie einen Code zur Erstellung von CUSUM-Diagrammen.*

2.2.2 Mehrdimensionale Skalierungs- und Korrespondenzanalyse-Biplots

Mit Hilfe der Verfahren der ordinalen mehrdimensionalen Skalierung kann, wie durch die Hauptkomponentenanalyse eine Objektkonfiguration erzeugt werden. Ein fundamentaler Unterschied zwischen Hauptkomponentenanalyse und ordinaler mehrdimensionaler Skalierung ist jedoch die Tatsache, daß die ordinale mehrdimensionale Skalierung von einer (n x n) Distanzmatrix **D** - die allerdings auch aus einer (n x p) Ausgangsmatrix **X** gebildet werden kann - ausgeht und nicht von der Datenmatrix **X** direkt[21]. Das heißt der q-dimensionale, durch die aus **D** berechnete Koordinatenmatrix $\mathbf{X}^*$ bestimmte Raum L, ist im Fall der mehrdimensionalen Skalierung kein Unterrraum des p-dimensionalen Raumes R der Matrix **X**. Von daher ist die ordinale mehrdimensionale Skalierung keine Projektions-, sondern eine Optimierungsmethode, die die, durch $\mathbf{X}^*$ definierten Objektdistanzen - auf iterativem Wege - möglichst nah an die tatsächlichen Objektdistanzen annähert. Die rechnerische Interpolation kann daher auch nur auf iterativem Weg erfolgen. Eine einfache graphische Interpolation im Sinne der Vektorsummenmethode basierend auf $\mathbf{X}^*$ ist aus diesem Grund ebenfalls nicht möglich. Vielmehr ist nach einer Transformation $\mathbf{X}^*\mathbf{Q}$ zu suchen, die einer Projektion von **X** in q Dimensionen so nahe wie möglich ist. Ähnlich ist bei der Ermittlung der Prediktionsmarker vorzugehen (zu den Einzelheiten siehe GOWER & HAND, 1996).

Die multiple Korrespondenzanalyse kann als Variante der Hauptkomponentenanalyse mit nominal- und ordinalskalierten Variablen (an Stelle der intervall- und verhältnisskalierten Variablen) beschrieben werden, wenn sie als Hauptkomponentenanalyse der doppelt gewichteten Indikatormatrix verstanden wird. Die bivariate Korrespondenzanalyse ist dann der Sonderfall für p = 2. Wie in der Hauptkomponentenanalyse führt die Eigenwertzerlegung auch in der Korrespondenzanalyse zur Minimierung der Abweichungen der im dimensionserniedrigten Raum gefundenen Distanzen von den Ausgangsdistanzen, nur daß es sich in der Korrespondenzanalyse um Chi-Quadrat Distanzen und nicht um euklidische Distanzen handelt[22]. Da es sich in der multiplen Korrespondenzanalyse der Indikatormatrix allerdings um dichotomisierte, in der Regel ursprünglich nominal- oder ordinalskalierte Variablen handelt, ist die Darstellung der Variablen in multiplen Korrespondenzanalyse-Biplots in Form kontinuierlicher Achsen weniger interessant. Vielmehr ergeben sich für die Kategorien der Variablen entsprechende Kategorien-Stufen-Punkte

---

[21] An dieser Stelle wird nach wie vor von einer Distanzmatrix mit euklidischen Distanzen ausgegangen, das heißt es handelt sich bei **D** um eine Distanzmatrix euklidischer Distanzen. Werden nicht-euklidische, aber euklidisch-einbettbare Distanzen verwendet, sind nichtlineare Biplots zu berechnen (siehe 2.2.3).

[22] Andere Distanzmaße als die Chi Quadrat-Distanz sind natürlich auch in der Korrespondenzanalyse möglich. Die Wichtung der Ausprägungen umgekehrt proportional zur Häufigkeit ihres Eintreffens - wie sie durch die Chi-Quadrat-Distanz erfolgt - ist sicher nicht in jedem Fall sinnvoll (GREENACRE, 1990).

(‘category level points’, CLPs), die jeweils eine Ausprägung einer Variablen charakterisieren.

Eine Darstellung der Objekte als Objekt-Punkte in Normalkoordinaten und der Variablen in Form von CLPs in Standardkoordinaten führt zur Biplot-Repräsentation der multiplen Korrespondenzanalyse. Die graphische Interpolation kann in diesem Fall nach der Vektorsummenmethode erfolgen. Die Verbindung der ein Objekt definierenden CLPs ergibt ein Polygon, dessen Zentroid der Lage des gesuchten Objekts entspricht. Die graphische Prediktion erfolgt nach Bildung von Prediktionsregionen. Die Prediktionsregion eines CLP ist diejenige Region, deren entfernteste Punkte dem, die Region definierenden CLP, näher sind als einem anderen CLP. Der Übersichtlichkeit halber ist es in der Regel sinnvoll für die Variablen separat Graphiken mit den Objekten und den jeweiligen Variablen und ihren Prediktionsregionen zu erstellen. Grundsätze zur Bildung derartiger Prediktionsregionen sind GOWER, 1993, zu entnehmen.

2.2.3 Nichtlineare und generalisierte Biplots

Hauptkomponentenanalyse-Biplots basieren auf der Annahme des Vorliegen der euklidischen Distanz. Daneben existieren auch nicht-euklidischer Proximitätsmaße. GOWER & LEGENDRE, 1986, zeigen aber, daß eine Vielzahl nicht-euklidischer Proximitätsmaße, euklidisch einbettbar ist[23]. Euklidisch einbettbar bedeutet, daß für das gewählte Proximitätsmaß eine Darstellung im euklidischen Raum in der Art möglich ist, daß die (euklidisch einbettbaren) Distanzen d_{rt} der Distanzmatrix **D**, aus den Distanzen zwischen den - durch die Koordinaten der Matrix $\mathbf{X}^*$ im euklidischen Raum definierten - Punkten hergeleitet werden können.

Die Koordinatenmatrix $\mathbf{X}^*$ ist hierbei definiert als die Lösung einer Hauptkoordinatenanlyse einer Distanzmatrix, deren Elemente euklidisch-einbettbare Distanzen sind. Handelt es sich um euklidische Distanzen, so entstehen die bereits besprochenen linearen Hauptkomponentenanalyse-Biplots. Werden euklidisch-einbettbare Distanzen verwendet und einer Hauptkoordinatenanalyse unterzogen, ergeben sich für intervall- und verhältnisskalierte sowie ordinalskalierte Variablen nichtlineare Biplots (GOWER & HARDING, 1988, MEULMAN & HEISER, 1993).

Jede Variable wird im nichtlinearen Biplot durch eine nichtlineare, mit Markern versehene Bahn ('trajectory') dargestellt. Diese Bahn entsteht durch die Berechnung sogenannter Pseudoobjekte. Diese Pseudoobjekte stehen für Objekte mit dem Wert ν für Variable j und 0 für alle anderen Variablen. Nimmt ν die Werte $0,\pm1,\pm2,\ldots$, an, entsteht durch die Pseudoobjekte die Variablenbahn für Variable j. Die Bahnen aller Variablen laufen in einem Punkt ***O*** zusammen, nämlich bei $\nu = 0$. Im linearen Biplot fallen der Zentroid der Ausgangsmatrix **X** und der Schnittpunkt der Biplotachsen ***O*** in einem Punkt zusammen, beim nichtlinearen Biplot ist dies nicht der Fall, sondern der Zentroid der Matrix $\mathbf{X}^*$, die die Koordinaten für die Projektion der Objekte liefert, und ***O*** unterscheiden sich in der Regel. Der Koordinatenvektor $\mathbf{m}_{\mathbf{int}}^{\boldsymbol{\nu}}$ für ein Pseudoobjekt errechnet sich - bei Zutreffen der Additivitätsannahme[24] - durch
$\mathbf{m}_{\mathbf{int}}^{\boldsymbol{\nu}} = \mathbf{L}^{-1}\mathbf{X}^{*\prime}((\mathbf{d}_{\mathbf{n+1}}^{\boldsymbol{\nu}})^2 - (1/\mathrm{n})\mathbf{D1})$, mit der Diagonalmatrix **L** der Eigenwerte von $\mathbf{B} = \mathbf{X}^*\mathbf{X}^{*\prime}$, der durch eine Hauptkoordinatenanlyse von **D** gewonnenen Koordinatenmatrix $\mathbf{X}^*$, der Einsermatrix **1** und dem (1 x n) Vektor $(\mathbf{d}_{\mathbf{n+1}}^{\boldsymbol{\nu}})^2$ der quadrierten Distanzen des gewählten Proximitätsmaßes des Pseudoobjekts zu den übrigen Objekten. Die Koordinaten sind für sich verändernde Werte von ν zu berechnen. Da jedes weitere Pseudoobjekt auch eine weitere

---

[23] Eine (notwendige) Bedingung für die euklidische Einbettbarkeit einer Distanzmatrix **D** ist, daß die zentrierte Matrix $\mathbf{B} = (\mathbf{I} - \mathbf{1s}')\mathbf{D}(\mathbf{I} - \mathbf{s1}')$ positiv semidefinit ist (**I**, (n x n) Einheitsmatrix; **D**, (n x n) Distanzmatrix; **1**, (n x 1) Matrix mit Einsen; **s** = **1**/n (Format n x 1) (GOWER & LEGENDRE, 1986).

[24] Proximitätsmaße werden als additiv bezeichnet, wenn jede Variable unabhängig von den anderen Variablen einen Beitrag zum Proximitätsmaß liefert. Ein Proximitätsmaß wie die Mahalanobis-Distanz, die auch die Kovarianzen der Variablen untereinander berücksichtigt, ist also zum Beispiel in diesem Sinn nicht additiv.

Dimension definiert, ergeben sich die entsprechenden Koordinaten jedes Pseudoobjekts auch in einer weiteren, der sogenannten Residualdimension. Die Interpretation wird jedoch durch diese Residualdimensionen nicht beeinträchtigt.

Die graphische Interpolation kann wie im linearen Fall durch die Vektorsummenmethode erfolgen, ausgehend vom Schnittpunkt der Biplotbahnen, nicht vom Zentroid der Objektdarstellung. *Für die Erstellung nichtlinearer Biplots auf Grundlage eines beliebigen Distanzmaßes mit Interpolationsmarkern auf den Biplotbahnen liegt ein Genstat Code im Anhang Teil III vor.*

Die Konstruktion von Prediktionsmarkern ist ebenfalls möglich (Einzelheiten siehe GOWER & HAND, 1996). Die graphische Prediktion erfolgt als sogenannte zirkuläre Prediktion und zwar in der Art, daß ein Kreis vom Zentroid zum Objektpunkt gebildet wird. Die Stelle an der der so entstandene Kreis die Biplotbahn schneidet, ergibt der vorhergesagten Variablenwert. Im linearen Fall ergeben zirkuläre und (bereits besprochene) orthogonale Prediktion denselben Markerwert auf der Biplotachse. Einige weitere Anmerkungen zu nichtlinearen Biplotbahnen:

1. entspricht die Position eines Pseudoobjekts dem Wert eines Objekts der Ausgangsmatrix **X**, so wird dieser Punkt als Basispunkt bezeichnet (siehe unten);
2. die nichtlinearen Biplotbahnen sind endlich, begrenzt durch den größten und den kleinsten Basispunkt;
3. die Verläufe der Bahnen können ausgesprochen bizarr sein, bis hin zu einer Umkehrung der Richtung und des Schneidens eines unteren Abschnittes durch einen oberen Abschnitt. Überschneidungen sind allerdings ein durch die Dimensionserniedrigung hervorgerufener Artefakt und entstehen nicht bei Betrachtung der vollen Dimensionalität. Derartig verzerrte Biplotbahnen sind jedoch sowohl für die Prediktion als auch für die Interpolation unzuverlässig.

GOWER, 1995b und GOWER & HAND, 1996, formulieren eine allgemeine ‘Biplot-Theorie’, die die Biplot-Darstellung beliebiger, auch gemischter Datensätze berücksichtigt, das heißt lineare Biplotachsen, nichtlineare Biplotbahnen und CLPs in einer Darstellung vereinen. Vorstellbar ist zum Beispiel die Verwendung des allgemeinen Ähnlichkeitskoeffizienten, mit dessen Hilfe eine Proximitätsmatrix für Variablen beliebiger Skalenarten gebildet werden kann, die dann durch die Ergebnisse einer Hauptkoordinatenanlyse in wenigen Dimensionen visualisiert wird. Durch Anwendung entsprechender Formeln, auf die hier nicht im Einzelnen eingegangen werden soll, erhält man über den Weg der Pseudoobjekte die Koordinaten der Biplotachsen und -bahnen und die CLPs der qualitativen Variablen. Die bereits angesprochenen Basispunkte der qualitativen Variablen entsprechen den CLPs und sind daher in diesem Zusammenhang, anders als bei den nichtlinearen Biplots quantitativer Variablen, von besonderem Interesse.

Bis auf Parallelverschiebungen gleichen die Biplotachsen beziehungsweise die Biplotbahnen der quantitativen Variablen denen der linearen und nichtlinearen Biplots. Allerdings gibt es in der Regel

für die Bahnen keinen gemeinsamen Schnittpunkt $\boldsymbol{O}$.

Die Interpolation erfolgt nach der Vektorsummenmethode unter Verwendung aller Variablen, das heißt der Werte auf den Biplotachsen, den nichtlinearen Biplotbahnen und den einem Objekt entsprechenden CLPs, ausgehend vom Zentroid der Objektdarstellung, das heißt ausgehend vom Zentroid von $\mathbf{X}^*$. Die Prediktionen lassen sich als zirkuläre Prediktion durchführen. Für die qualitativen Variablen sind entsprechend Prediktionsregionen zu erstellen.

2.2.4 ‘Klassische’ Biplots

Biplots gehen ursprünglich zurück auf GABRIEL, 1971, und werden zum Beispiel von GABRIEL, 1981, GABRIEL & ODOROFF, 1986, oder auch GABRIEL, 1995a & 1995b, dargestellt und diskutiert. Die klassische Formulierung bedient sich der Eigenwertzerlegung der Datenmatrix **X**. Unterschieden wird zwischen dem CMP (column preserving), RMP (row preserving) und dem diagnostischen Biplot. Die Darstellung der ‘klassischen’ Biplots erfolgt vielfach als Punkt- und Pfeile-Plots, das heißt, die Zeilen (die Objekte) werden durch die Endpunkte der vom Ursprung ausgehenden Vektoren als Punkte, die Spalten (die Variablen) durch die vom Ursprung ausgehenden Vektoren in Form von Pfeilen dargestellt. Der diagnostische Biplot hat vor allem in der Modellwahl, zum Beispiel in der Regressionsanalyse eine Bedeutung (GABRIEL, 1981). Wichtige Merkmale der CMP- und RMP-Biplots als Repräsentation der Datenmatrix **X** sind (GABRIEL, 1995a):

1. der CMP-Biplot approximiert die Mahalanobis-Distanz zwischen den Objekten, der RMP-Biplot die euklidische Distanz;
2. die Länge der jeweiligen Variablenachsen im CMP-Biplot approximiert die Standardabweichnug der Variablen. Ist diese bekannt, so ist die Länge des Variablenvektors ein guter Anhaltspunkt für die Güte der Repräsentation dieser Variablen im Biplot. Besonders prägnant wird dies im Fall standardisierter Variablen mit Standardabweichung = 1, da dann ein vom Ursprung ausgehender Kreis mit Radius 1 das Maximum der Standardabweichnug der Variablen vorgibt und schnell sichtbar wird, wie gut oder wie schlecht eine Variable repräsentiert ist;
3. der Kosinus des Winkels zwischen zwei Variablenachsen approximiert im CMP-Biplot die Korrelation zwischen zwei Variablen;
4. die Elemente der Ausgangsmatrix **X** werden sowohl im CMP- als auch im RMP-Biplot durch das innere Produkt eines Spalten- und eines Zeilenvektors approximiert. Dies ermöglicht zum Beispiel Schlußfolgerungen über die Bedeutung bestimmter Variablen bei ausgewählten Objekte oder über die tatsächlichen Werte bestimmter Objekte bei bestimmten Variablen und damit also auch über die Unterschiede von Objekten bei den Variablen. Diese Eigenschaft liefert letztlich auch die Grundlage der bereits besprochenen Interpolations- und Prediktionseigenschaften der Biplots.

Bei Vorliegen qualitativer Daten in Form einer Kontingenztafel oder einer Indikatormatrix schlägt GABRIEL, 1995a und 1995b, als Alternative zu Abbildungen, die durch die Korrespondenzanalyse gewonnenen werden, die Verwendung separater Reihenprofil- beziehungsweise Spaltenprofil-Biplots vor. Ob und inwieweit die getrennte Darstellung als Reihenprofil- beziehungsweise Spaltenprofil-Biplot der traditionellen Darstellung der bivariaten Korrespondenzanalyse überlegen ist, soll an dieser Stelle nicht vertieft werden (siehe aber dazu zum Beispiel die Diskussionen bei

GABRIEL, 1995b oder GREENACRE, 1993).

2.3 Analyse gruppierter Daten

In vielen Fällen der Datenanalyse liegen die Daten in der einen oder anderen Art gruppiert vor. Im Kontext dieser Arbeit sind Faktoren, die diese Gruppen bestimmen zum Beispiel verschiedene Variablensets oder Erhebungsjahre. Gruppierte Daten können variablen- oder objektorientiert analysiert werden. Variablenorientiert heißt in diesem Zusammenhang, daß die Frage gestellt wird, ob die Variabilitätsstruktur in den Gruppen als gleich oder als unterschiedlich angesehen werden kann, ob also zum Beispiel die Ausrichtung der Achsen, das heißt der ersten, zweiten, dritten und so weiter Hauptkomponente, in etwa gleich ist oder nicht. Werden im gruppierten Fall Hauptkomponentenanalysen für die einzelnen Gruppen getrennt durchgeführt, stellt sich die Frage, in wie weit sich die für die jeweiligen Gruppen ermittelten Eigenwerte und Eigenvektoren ähnlich beziehungsweise unähnlich sind. Diese Fragestellung kann mit Hilfe gemeinsamer Hauptkomponentenmodelle untersucht werden (siehe 2.3.1). Objektorientierte Ansätze fragen demgegenüber danach, ob die Objekte in aus verschiedenen Analysen abgeleiteten Konfigurationen, an derselben Stelle liegen oder stark voneinander entfernt sind, ob also zum Beispiel zwischen der Konfiguration der Punkte der Objekte im Koordinatensystem einer ersten Gruppe eine gute oder schlechte Übereinstimmung mit der Konfiguration der Punkte der Objekte im Koordinatensystem einer zweiten Gruppe besteht. Dieser Fragestellung kann mit Hilfe der Prokrustes-Analyse nachgegangen werden (siehe 2.3.2). Alternative, objektorientierte Methoden sind die gewichtete mehrdimensionale Skalierung und die kanonische Variablenanalyse. Eine weitere variablenbezogene Methode ist die nichtlineare kanonische Analyse. Die genannten drei Verfahren werden in Kapitel 2.3.3, aufgrund ihres geringen Gewichts in der vorliegenden Arbeit allerdings nur kurz, angesprochen.

2.3.1 Gemeinsame Hauptkomponentenmodelle

2.3.1.1 Gemeinsame Hauptkomponenten

Wenn die Daten in Form getrennter Stichproben gruppenweise strukturiert vorliegen, ist die Frage zu stellen, ob sich die Hauptkomponentenanalysen in den einzelnen Gruppen einander ähneln oder stark von einander abweichen. Sind sie sich sehr ähnlich, kann die Beschreibung der Gruppen mit Hilfe des gemeinsamen Hauptkomponentenmodells erfolgen (FLURY, 1984 und 1988, FLURY & RIEDWYL, 1988). Ähnlichkeit ist im variablenorientierten Ansatz dieses Modells so zu verstehen, daß die Ausrichtung der Achsen (nicht notwendigerweise ihre relative Bedeutung und Größe), in allen Gruppen annähernd gleich ist, die einzelnen Gruppen sich demnach durch ein gemeinsames Achsensystem (eine gemeinsame Transformation) ohne einen erheblichen Informationsverlust beschreiben lassen.

Die Vorteile dieses Modells sind zum einen die Vereinfachung der Ergebnisdarstellung, wenn nur

eine gemeinsame Hauptkomponentenanalyse, an Stelle einer Vielzahl separater Hauptkomponentenanalysen, präsentiert werden muß. Darüber hinaus werden bei Verwendung nur einiger Hauptkomponenten im reduzierten, gemeinsamen Modell, für alle Gruppen die gleichen Hauptkomponenten verworfen, und somit die Gruppen im gleichen, reduzierten Variablenraum beschrieben.

Allerdings sind die gemeinsamen Hauptkomponenten in der Regel nicht wie die ursprünglichen Hauptkomponenten unkorreliert. Zudem ist die Anwendung des gemeinsamen Hauptkomponenten-Modells bislang nur bei Verwendung der Kovarianzmatrix ausreichend entwickelt. Abweichungen von der Multinormalverteilung sowie das Vorliegen von Ausreißern können die Schätzmethoden des gemeinsamen Hauptkomponentenmodells stark beeinflussen.

Das gemeinsame Hauptkomponentenmodell ist nur ein Modell in einer Hierarchie von Modellen zur Beschreibung der Beziehung der Kovarianzmatrizen gruppierter Daten. Folgende Modelle lassen sich voneinander abgrenzen:

1. die Kovarianzmatrizen der Gruppen sind gleich;
2. die Kovarianzmatrizen der Gruppen sind proportional zueinander;
3. die Kovarianzmatrizen der Gruppen lassen sich durch das gemeinsame Hauptkomponenten-Modell beschreiben;
4. die Kovarianzmatrizen der Gruppen lassen sich durch das partielle, gemeinsame Hauptkomponentenmodell beschreiben;
5. die Kovarianzmatrizen der Gruppen sind voneinander verschieden und nicht proportional zueinander.

Die Fälle 1. und 5. können durch bekannte Testverfahren auf Gleichheit der Kovarianzmatrizen bearbeitet werden (MORRISON, 1990). Die Fälle 2., 3., und 4. werden von FLURY, 1984, und FLURY & RIEDWYL, 1988, behandelt. KRZANOWSKI, 1984, liefert dafür ein approximatives Vorgehen. Aufgrund der weitreichenden Modellannahmen und der Notwendigkeit der Verwendung der Kovarianzmatrix zur Berechnung der Maximum Likelihood Schätzer ist dieses gemeinsame Hauptkomponentenmodell für die vorliegende Arbeit nicht geeignet.

2.3.1.2 Gruppenanalysemodell

Das hier beschriebene Gruppenanalysemodell geht zurück auf KRZANOWSKI, 1979 und 1988a. Im Prinzip verfolgt es dieselben Ziele wie 2.3.1.1. Es geht also um die Frage, in wie weit die Variabilitätsstruktur verschiedener Gruppen durch ein gemeinsames Hauptkomponentenmodell dargestellt werden kann. Die Grundsätze lassen sich wie folgt skizzieren:

für zwei Gruppen A und B liegen die Koeffizienten der Hauptkomponentenanalysen (die Eigenvektoren) als l_{jA} und m_{jB} vor ($j_A = 1 \ldots q_A$), ($j_B = 1 \ldots q_B$), ($q_A \leq p$ und $q_B \leq p$). **L** und **M** sind die (q_A x p) und (q_B x p) Matrizen der Eigenvektoren der Hauptkomponentenanalysen von

A und B. Wenn nun **N** = **LM'ML'** definiert wird, gilt:

1. der kleinste Winkel zwischen einem beliebigen Vektor der ersten q Hauptkomponenten von A und dem am parallesten gelegenen Vektor der ersten q Hauptkomponenten von B, ist $\alpha_1 = \cos^{-1}\sqrt{l_1}$, wobei l_1 der größte Eigenwert von **N** ist.
2. Wenn l_j der j-größte Eigenwert von **N**, $\mathbf{a_j}$ der zugehörige Eigenvektor und $\mathbf{b_j} = \mathbf{L'a_j}$ sind, dann sind $\mathbf{b_1} \ldots \mathbf{b_q}$ orthogonale Vektoren im Raum von A und $\mathbf{M'Mb_1} \ldots \mathbf{M'Mb_q}$ orthogonale Vektoren im Raum B, und der Winkel zwischen dem j-ten Paar $\mathbf{b_j}$ und $\mathbf{M'Mb_j}$ ist $\alpha_j = \cos^{-1}\sqrt{l_j}$.

Die Summe der Eigenwerte von **N** ist gleich der Quadratsumme der Kosinen der Winkel der Hauptkomponenten zwischen A und B. Sind die Achsen der beiden Gruppen völlig übereinstimmend, so nimmt diese Summe den Wert g (= Anzahl der Gruppen), sind sie orthogonal, den Wert Null, an. Nun können also die Ähnlichkeiten zwischen A und B durch die Vektorpaare $\mathbf{b_j}$ und $\mathbf{M'Mb_j}$ dargestellt werden. l_j repräsentiert dann den Beitrag des j-ten Paars zur Gesamtvariabilität. Die Linie im Raum A, die dem Raum B am nächsten liegt ist gegeben durch $\mathbf{b_1}$; $\mathbf{b_1}$ liegt am nächsten zu $\mathbf{M'Mb_1}$ in B und der Winkel zwischen ihnen ist $\alpha_1 = \cos^{-1}\sqrt{l_1}$. Die Ebene im Raum A, die dem Raum B am nächsten liegt, ist definiert durch die Vektoren $\mathbf{b_1}$ und $\mathbf{b_2}$, und entsprechend definieren die Paare ($\mathbf{b_1}$, $\mathbf{b_2}$) und ($\mathbf{M'Mb_1}$, $\mathbf{M'Mb_2}$) die sich von A und B am nächsten gelegenen Ebenen mit den 'kritischen' Winkeln $\cos^{-1}\sqrt{l_1}$ und $\cos^{-1}\sqrt{l_2}$. Diese Aufteilung läßt sich nun fortführen für alle von A und B gemeinsam beschriebenen Dimensionen q. Die 'kritischen' Winkel geben Aufschluß darüber, wie gut oder wie schlecht die Übereinstimmung der Achsen im q-dimensionalen Raum ist. Eine völlige Übereinstimmung führt zu 'kritischen' Winkeln mit dem Wert Null; die Eigenwerte von **N** sind dementsprechend dann alle gleich 1. $\mathbf{b_1} \ldots \mathbf{b_q}$ schließlich stellen in diesem Fall die Koeffizienten des gemeinsamen q-dimensionalen Raums von A und B dar. Entsprechen sich die Achsen nicht, so ist der mittlere Vektor von $\mathbf{b_j}$ und $\mathbf{M'Mb_j}$ definiert durch $\mathbf{c_j} = \left[2\left(1+\sqrt{l_j}\right)\right]^{-1/2}\left(\mathbf{I} + 1/\sqrt{l_j}\mathbf{M'M}\right)\mathbf{b_j}$ und die $\mathbf{c_1} \ldots \mathbf{c_q}$ definieren die mittleren Komponenten der Dimensionen von A und B.

Wenn A und B durch eine unterschiedliche Anzahl von Hauptkomponenten charakterisiert sind, und q_A die Anzahl der Hauptkomponenten von A und q_B die Anzahl der Hauptkomponenten von B sind, ist q = min (q_A, q_B). **N** hat q von Null verschiedene Eigenwerte und der Vergleich von A und B erfolgt auf der Basis eben dieser von Null verschiedenen Eigenwerte.

Die Ausweitung des Konzepts auf mehr als zwei Gruppen kann folgendermaßen verdeutlicht werden:

$\mathbf{L_t}$ ist die (q x p) Matrix mit den Eigenvektoren der Hauptkomponentenanalyse der Gruppe t. **b** ist

ein Vektor im Raum der Ausgangsvariablen, und δ_t der Winkel zwischen **b** und der am nächsten gelegenen Achse im durch die q, von Gruppe t definierten, Hauptkomponenten. Der Wert von **b** der $V = \sum_{t=1}^{g} \cos^2 \delta_t$ maximiert, ist gegeben durch den Eigenvektor $\mathbf{b_1}$, der wiederum korrespondiert zum größten Eigenwert l_1 von $\mathbf{H} = \sum_{t=1}^{g} \mathbf{L_t'L_t}$. Ein Maß für die Abweichung von $\mathbf{b_1}$ vom durch Gruppe t definierten Raum ist $\delta_{t1} = \cos^{-1}\left[\left(\mathbf{b_1'L_t'L_tb_j}\right)^{1/2}\right]$. Die Eigenwert/Eigenvektor Zerlegung von **H** führt zu den Vektoren $\mathbf{b_1} \ldots \mathbf{b_q}$, die einen Unterraum des Ausgangsdatenraums beschreiben, der allen Gruppen gleichzeitig so nahe wie möglich ist. Als Maß für die Abweichung der Gruppen von diesen Vektoren ist δ_{tj} definiert als $\delta_{tj} = \cos^{-1}\left[\left(\mathbf{b_j'L_t'L_tb_j}\right)^{1/2}\right]$ mit t = 1 ... g und j = 1 ... q (q ≤ p). Dann ergibt sich $V = \sum_{t=1}^{g}\sum_{j=1}^{q} \cos^2 \delta_{tj}$. Das Maximum von V liegt bei g, da dann völlige Übereinstimmung der Hauptkomponenten vorliegt und alle δ_{tj} gleich Null sind. *Für das Gruppenanlysemodell liegt ein Genstat Code im Anhang Teil III vor.*

2.3.1.3 Gamma-q-q-Plots

Eine graphische Methode zum Vergleich der Eigenvektoren der Hauptkomponentenanalysen verschiedener Gruppen wird von KERAMIDAS et al., 1987, vorgestellt. Beschrieben ist sie ausschließlich für den Fall, daß als Ausgangspunkt die Kovarianzmatrix verwendet wird, obwohl KERAMIDAS et al., 1987, eine Übertragung auf den Fall Ausgangspunkt Korrelationsmatrix für denkbar halten. Wichtig für den Einsatz der Methode sind:

1. die Anzahl der zu vergleichenden Gruppen sollte möglichst groß sein (t ≥ 10);
2. die Eigenwerte der Hauptkomponenten sollten sich deutlich von einander abheben;
3. ein Vergleichsmaßstab in Form eines a priori festgelegten Eigenvektors ($\boldsymbol{\xi}_\mathbf{j}$) beziehungsweise aus den Daten bestimmten 'typischen' Eigenvektors ($\mathbf{b_j}$) muß vorgegeben werden. Der 'typische' Eigenvektor ist dabei derjenige Eigenvektor, der die Winkel zwischen sich selbst und den entsprechenden Eigenvektoren der Gruppen minimiert und kann errechnet werden als der Eigenvektor des größten Eigenwertes von $\mathbf{H} = \sum_{t=1}^{g} \mathbf{L_t'L_t}$ (siehe 2.3.1.2).

Um festzustellen, ob sich die Eigenwerte gut von einander abheben, können Boxplots der Eigenwerte der Gruppen hilfreich sein. Die Eigenwerte aller Gruppen werden in Form von Boxplots so dargestellt, daß auf der x-Achse die laufende Nummer der Eigenwerte, auf der y-Achse die Boxplots der Eigenwerte abgetragen werden. Aus diesen Plots wird erkennbar, wie groß die Unterschiede zwischen den Eigenwerten aller Gruppen sind, und welche Überschneidung zwischen den Eigenwerten aller Gruppen vorliegt. Sie erlauben also einen gleichzeitigen, groben Einblick in

alle Eigenwerte aller Gruppen.

Der notwendige Vergleichsmaßstab bei Betrachtung einer Hauptkomponente wird durch die euklidische Distanz für den a priori Vektor $\boldsymbol{\xi}_j$ (beziehungsweise $\mathbf{b}_j$ an Stelle von $\boldsymbol{\xi}_j$) vom Beobachtungsvektor durch $\delta_{tj}^2 = \min\left(\boldsymbol{\xi}_j - \mathbf{a}_{tj}\right)'\left(\boldsymbol{\xi}_j - \mathbf{a}_{tj}\right), \left(\boldsymbol{\xi}_j + \mathbf{a}_{tj}\right)'\left(\boldsymbol{\xi}_j + \mathbf{a}_{tj}\right)$ errechnet, und zwar für Gruppe t (t = 1 ... g) und die Koeffizienten des Eigenvektors (der Hauptkomponente) $\mathbf{a}_j$ (j = 1 ... p). Da diese Distanzen gut durch eine Gamma-Verteilung approximiert werden können, kann ein Gamma-q-q-Plot erstellt werden, bei dem auf der x-Achse die Gamma-Quantile, auf der y-Achse die geordneten δ_{tj}^2-Werte aufgetragen werden. Die Gamma Quantile werden ermittelt nach vorheriger Schätzung der Form-(η) und Größe-(λ) Parameter der Gamma-Verteilung aus den ermittelten Distanzen. Der q-q-Plot zeigt die Gamma Quantile auf Grund der geschätzten Parameter und die für jede Gruppe kleinsten quadrierten euklidischen Distanzen, die der Eigenvektor der Hauptkomponente zum 'typischen' oder a priori Eigenvektor hat. Eine deutliche Abweichung des q-q-Plots von der Linearität weist für die Gruppen, die diese Abweichung verursachen, auf einen vom a priori beziehungsweise 'typischen' Eigenvektor deutlich abweichenden Eigenvektor und damit bei diesen Gruppen auf eine vom Vergleichsmaßstab abweichende Kovarianzstruktur hin. Beim Vergleich von mehr als einer Hauptkomponente wird δ_{tj}^2 zu $\Delta = \sum_{t=1}^{g} \sum_{j=1}^{q} 2\left(1 - \delta_{tj}\right)$. Wie bereits erwähnt ist die vorgeschlagene Methode nur bei einer sehr großen Gruppenanzahl, vorzugsweise bei Verwendung der Kovarianzmatrix anwendbar. *Der Genstat Code zur Erstellung der Gamma-q-q-Plots und Eigenwerte-Boxplots liegt im Anhang Teil III vor.*

2.3.2 Prokrustes-Analyse

Die Prokrustes-Analyse dient zum objektorientierten Vergleich zweier oder mehrerer Konfigurationen. Mit Konfiguration ist hier die durch die Variablenwerte bestimmte Lage der Objekte im p-dimensionalen Raum gemeint. Nicht die Übereinstimmung der Werte, das heißt die absolute Lage der Objekte im Koordinatensystem, bildet dabei den Maßstab für die Beurteilung der Übereinstimmung von Konfigurationen, sondern die relative Lage der Objekte zueinander in den Koordinatensystemen unterschiedlicher Konfigurationen.

Unterschiedliche Konfigurationen derselben Objekte können entstehen durch:

1. Analyse unterschiedlicher Variablen derselben Objekte;
2. Analyse unterschiedlicher Gruppen (als Variablengruppierung, wenn zum Beispiel dieselben Variablen in verschiedene Jahren bestimmt werden);
3. Analyse von Wiederholungen;
4. Analyse derselben Daten mit unterschiedlichen Methoden (Hauptkomponentenanalyse, mehrdimensionale Skalierung, Korrespondenzanalyse, verschiedene Proximitätsmaße).

Zu unterscheiden ist zwischen der einfachen Prokrustes-Analyse für den paarweisen Vergleich von zwei Konfigurationen und der generalisierten Prokrustes-Analyse für den gleichzeitigen Vergleich von mehr als zwei Konfigurationen. Da nicht die absolute Lage der Objekte im Koordinatensystem für die Prokrustes-Analyse von Bedeutung ist, sondern die relative Lage der Objekte zueinander, ist es sinnvoll, verschiedene Datenmanipulationen durchzuführen, die dafür sorgen, daß die Übereinstimmung der Koordinaten der verschiedenen Konfigurationen zu gut wie irgend möglich ist; das heißt, es sind Transformationen durchzuführen, die die inneren Beziehungen der jeweiligen Konfigurationen bewahren. Erst dann ist ein Maß für die Übereinstimmung der Konfigurationen im Sinne der Prokrustes-Analyse zu berechnen. Die genannten Datenmanipulationen umfassen:

1. Translation, das heißt eine feste Lageveränderung aller Punkte um eine gemeinsame Entfernung in einer gemeinsamen Richtung;
2. Rotation, das heißt eine feste Lageveränderung aller Punkte um einen gemeinsamen, konstanten Winkel, die die Distanz eines jeden Punktes vom Zentroid unberührt läßt. Eine Reflexion (Spiegelung) kann als Form der Rotation verstanden werden; und
3. Dilation, das heißt ein Strecken beziehungsweise Stauchen aller Punkte durch eine Konstante an einer Linie vom Objektpunkt vom (beziehungsweise zum) Zentroid.

Die Variablen der zu vergleichenden Ausgangsmatrizen sind möglicherweise vor der Prokrustes-Analyse zu standardisieren. Besitzen unterschiedliche Matrizen eine unterschiedliche Anzahl an Spalten, so gilt, daß die Matrizen, deren Variablenzahl < p_{max} ist, durch Nullspalten ergänzt

werden. Die Prokrustes-Analyse hat starke Impulse von GOWER, 1975 und 1995a, erhalten und wird in der Folge im Sinne dieser Referenzen dargestellt.

Liegen zwei Konfigurationen in Form der (n x p) Matrizen **X** und $\mathbf{X}^*$ vor, mit den Elementen x_{ij} und x^*_{ij}, so ist M^2 als Maß für die Abweichung der einen Konfiguration von der anderen Konfiguration wie folgt definiert: $M^2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{p}(x_{ij} - x^*_{ij})^2\right]$. Vor Berechnung dieser Maßzahl sind die oben angesprochenen Transformationen durchzuführen.

Die Translation wird erreicht durch die Mittelwertszentrierung der Ausgangsmatrizen **X** und $\mathbf{X}^*$. Sie führt dazu, daß der Zentroid von **X**, $\mathbf{G_X}$ gleich dem Mittelwertsvektor von $\mathbf{X}^*$, $\mathbf{G_{X^*}}$ ist und es gilt: $\mathbf{G_X} = \mathbf{G_{X^*}} = \mathbf{0}$, das heißt, beide Konfigurationen haben denselben Zentroid, gelegt am Ursprung. Mögliche Unterschiede zwischen den Mittelwertsvektoren der Konfigurationen werden durch die Translation also entfernt. Sind diese von Interesse, kann vor der Translation eine multivariate Varianzanalyse durchgeführt werden, die jedoch in dieser Arbeit nicht betrachtet wird.

Rotation und Dilation werden nach Translation beider Matrizen derart durchgeführt, daß eine Matrix als fix (zum Beispiel **X**) die andere Matrix als beweglich (zum Beispiel $\mathbf{X}^*$) angenommen wird. Da die Dilation nicht symmetrisch ist (das heißt der Faktor c, der M^2 minimiert, bei Skalenveränderung von $\mathbf{X}^*$ gegeben **X**, ist nicht notwendigerweise gleich dem Faktor c^*, der M^2 minimiert, bei Skalenveränderung von **X** gegeben $\mathbf{X}^*$), werden die Variablen in der Regel so standardisiert, daß gilt $\text{spur}(\mathbf{XX}') = \text{spur}(\mathbf{X}^*\mathbf{X}^{*\prime})$. Als Konsequenz ergeben sich $c^2 + M^2_{\text{minimum}} = 1$ und $c = c^*$. Werden mehrere Konfigurationen paarweise miteinander verglichen, so können die jeweiligen M^2-Werte der Paarvergleiche als Proximitätsmaß betrachtet werden und zum Beispiel einer Hauptkoordinatenanlyse unterzogen werden, die dann wiederum eine Konfiguration erzeugt, die die Lage der unterschiedlichen, paarweise miteinander verglichenen Konfigurationen aufzeigt.

Die Generalisierung der einfachen Prokrustes-Analyse für den gleichzeitigen Vergleich von mehr als zwei Konfigurationen erfolgt (nach Mittelwertszentrierung) im Gegensatz zur einfachen Prokrustes-Analyse auf iterativem Weg, da zur Ermittlung der Dilationsfaktoren und der Rotationsmatrizen, die die Abweichungen minimieren, die mittlere Endkonfiguration bekannt sein muß. Da sie das natürlich nicht ist, kann man sich ihr nur bis zu einem gewissen Konvergenzkriterium nähern. Die mittlere Konfiguration nach Abschluß der Datenmanipulationen im Rahmen der Prokrustes-Analyse, wird als Konsens-Konfiguration bezeichnet.

Um die unterschiedlichen Begriffe in der Prokrustes-Analyse noch einmal zu verdeutlichen und die Einbindung der Ergebnisse einer Prokrustes-Analyse in ein varianzanalytisches Schema aufzuzeigen, sei nach GOWER, 1995a, folgendermaßen definiert:

es liegen g (n x p_t) Matrizen $\mathbf{X_t}$ (t = 1 ... g) vor mit $p = p_{max}(p_1, p_2, \ldots, p_g)$ Dimensionen. Das i-te Objekt der t-ten Konfiguration belegt den Punkt A_{it} mit den Koordinaten $(x_{i1t}, x_{i2t}, \ldots, x_{ipt})$. Der Zentroid der t-ten Konfiguration ist G_t, der Zentroid aller Punkte G.

Handelt es sich bei den Variablen um dieselben Variablen in allen Konfigurationen, so sind die Mittelwertsvektor-unterschiede durch eine multivariate Varianzanalyse analysierbar nach dem Modell

$$\sum_{t=1}^{g}\sum_{i=1}^{n}(G - A_{it}) = n\sum_{t=1}^{g}(G - G_t)^2 + \sum_{t=1}^{g}\sum_{i=1}^{n}(G_t - A_{it})^2$$ (Total = Translation + Residuen).

Duch die Translation werden die Mittelwertsvektorunterschiede eliminiert und die Konfigurationen zu einem gemeinsamen Ursprung *O* überführt. Als Varianzanalyse-Modell läßt sich dann schreiben:

$$\sum_{t=1}^{g}\sum_{i=1}^{n}(O - A_{it})^2 = g\sum_{i=1}^{n}(F_i - O)^2 + \sum_{t=1}^{g}\sum_{i=1}^{n}(F_i - A_{it})^2$$ (Total = Konfigurationseffekt + individuelle Differenzen (Residuen)), wobei F_i der Zentroid von Objekt i für alle Konfigurationen t ist. Die Gesamtstreuung um den Ursprung wird also aufgeteilt in zwei Orientierungseffekte. Zum einen die Abweichung der Zentroide der Konfigurationen der Objekte vom Ursprung (erster Orientierungseffekt, Konfigurationseffekt), zum anderen die Abweichung der Koordinaten der Objekte vom jeweiligen Konfigurationszentroid (zweiter Orientierungseffekt, individuelle Differenzen). Der erste Orientierungseffekt kann nur objektbezogen, der zweite Orientierungseffekt sowohl objekt- als auch konfigurationsbezogen ermittelt werden. So gilt also objektbezogen zum Beispiel für Objekt 1

$$\sum_{t=1}^{g}(O - A_{1t})^2 = g(F_1 - O)^2 + \sum_{t=1}^{g}(F_1 - A_{1t})^2$$ (Total = Konfigurationseffekt + individuelle Differenzen). Konfigurationsbezogen errechnet sich das Residuum zum Beispiel für Konfiguration 1 durch $\sum_{i=1}^{n}(F_i - A_{i1})^2$. Ein Konfigurationseffekt kann bei der konfigurationsbezogenen Betrachtung natürlich nicht berechnet werden.

Ein großes Objektresiduum weist darauf hin, daß Objekte in Konfigurationen stark voneinander abweichen. Ein großes Konfigurationsresiduum weist darauf hin, daß die Unterschiede der Konfigurationen von der Konsens-Konfiguration erheblich sind. Ein starker Translationseffekt ist ein Indiz für Unterschiede der Mittelwertsvektoren der Ausgangsmatrizen. Unterschiede bei den Dilationsfaktoren deuten auf Unterschiede in der absoluten Skala der Variablen der verschiedenen Konfigurationen hin (zum Beispiel durch unterschiedlich starke Variabilität in den verschiedenen Konfigurationen).

2.3.3 Gewichtete mehrdimensionale Skalierung, kanonische Variablenanalyse und nichtlineare kanonische Analyse

2.3.3.1 Gewichtete mehrdimensionale Skalierung

Ein der Prokrustes-Analyse vergleichbares Verfahren stellt die gewichtete mehrdimensionale Skalierung dar, die auf eine Arbeit von CAROLL & CHANG, 1970, zurückgeht. Sie wird häufig auch als individuelle Differenzskalierung bezeichnet. Neben den Koordinaten der Objekte in q Dimensionen erzeugt die gewichtete mehrdimensionale Skalierung in Form von Gewichtungswerten Maßzahlen für die Bedeutung der jeweiligen Dimension für die verschiedenen Gruppen und erlaubt somit einen Vergleich derselben. Die gewichtete mehrdimensionale Skalierung ist mit Hilfe eines iterativen Algorithmus zu lösen, ist somit der ordinalen mehrdimensionale Skalierung nahe. Sie kann aber im wesentlichen als Generalisierung der Hauptkoordinatenanalyse für mehr als eine Proximitätsmatrix angesehen werden und beinhaltet demnach auch Aspekte dieses Verfahrens. Der gewichteten mehrdimensionalen Skalierung liegt der Gedanke zugrunde, daß sich alle Gruppen durch ein gemeinsames Achsensystem beschreiben lassen und die Unterschiede zwischen den Gruppen durch das Gewicht, das die einzelnen Gruppen den jeweiligen gemeinsamen Achsen zuordnen, gegeneinander abgegrenzt werden können und daß, durch Verwendung von gruppenspezifischen Gewichtungswerten, aus dem gemeinsamen q-dimensionalen Raum ($q \leq n$) ein jeweils gruppenspezifischer Raum (mit $\leq q$ Dimensionen) errechnet werden kann. Einige Kennwerte der gewichteten mehrdimensionalen Skalierung sind (CARROL, 1972, SCHIFFMANN et al., 1981):

1. Die Koordinaten der Gesamtkonfiguration aller Gruppen in q Dimensionen; da dieser Konfiguration eine Iteration zugrunde liegt, muß sie bei Veränderung der Anzahl q für alle Dimensionen neu errechnet werden.
2. Die Gewichtungswerte w_{tj} je Gruppe und Dimension; ist $w_{tj} = 1$, so ist die Dimension j für Gruppe t gleich der Dimension j der Gesamtkonfiguration. Ist $w_{tj} <$ 1, ist die Dimension j der Gruppe t im Vergleich zur Gesamtkonfiguration gestaucht, ist $w_{tj} > 1$, ist sie gestreckt. Je größer beziehungsweise kleiner w_{tj} ist, desto größer beziehungsweise kleiner ist die Bedeutung von Dimension j für Gruppe t. Ähnliche oder gleiche Gewichtungswerte zweier Gruppen deuten auf die Ähnlichkeit der Konfigurationen der Gruppen hin und umgekehrt.
3. Werden die Gewichtungswerte als vom Ursprung ausgehende Vektoren betrachtet, so ist der Winkel zweier Vektoren ein Maß für die Ähnlichkeit der Gruppen[25]. Die

[25]Dieser Gedanke kann noch vertieft werden durch die sogenannte Winkel-Varianzanalyse (analysis of angular variation), wenn sich die Gruppen inhaltlich sinnvoll in Obergruppen einteilen lassen (JONES, 1983, SCHIFFMANN et al., 1981).

Länge des jeweiligen Vektors ist ein Maß für die Anpassungsgüte von jeweiliger Gruppenkonfiguration und Gesamtkonfiguration. Wenn O_t die Länge des gruppenspezifischen Vektors der Gruppe t ist gilt: $O_t = \sqrt{\sum_{j=1}^{q} w_{tj}^2}$ und $O_t^2 = \sum_{j=1}^{q} w_{tj}^2$, wobei O_t^2 ein Maß für den Anteil der Gesamtvariabilität ist, der durch das gewählte Modell repräsentiert wird.

Abschließend soll darauf hingewiesen werden, daß die Ergebnisse der gewichteten mehrdimensionalen Skalierung keiner Rotation unterzogen werden dürfen und daß negative Gewichtungswerte zwar grundsätzlich möglich, in der Anwendung jedoch sehr selten sind.

2.3.3.2 Kanonische Variablenanalyse

Die kanonische Variablenanalyse geht von einer, am Ursprung zentrierten, in g Gruppen aufgeteilten (n x p) Datenmatrix **X** aus. Gesucht wird nach der Linearkombination der p Variablen, die das Verhältnis von der SSP (Sums of Squares and Products)-Matrix **B** (Between, zwischen den Gruppen) zu der SSP-Matrix **W** (Within, innerhalb der Grupppen) maximiert, und damit eine Funktion erzeugt, die die vorhandenen Gruppen im Sinne einer kleinsten-Quadrate-Lösung, optimal zu trennen in der Lage ist[26]. Diese Lösung wird erzielt durch die Eigenwertanalyse von $\mathbf{W}^{-1}\mathbf{B}$. Der mit dem ersten Eigenwert von $\mathbf{W}^{-1}\mathbf{B}$ assoziierte Eigenvektor bestimmt die Richtung im p-dimensionalen Raum an dem die Variabilität zwischen den Gruppen am größten ist, im Vergleich zur Variablität innerhalb der Gruppen. In Analogie zur Hauptkomponentenanalyse bestimmen die folgenden Eigenvektoren die nächstwichtigen Dimensionen im Sinne der Maximierung des Verhältnisses von Between-Streuung zu Within-Streuung. Sowohl die Objekte als auch die Gruppenmittelwertsvektoren lassen sich mit Hilfe der Eigenvektoren in den Raum der kanonischen Variablen projizieren und damit in wenigen Dimensionen graphisch abbilden (CHATFIELD & COLLINS, 1980). Der unterschiedlichen Variablität der einzelnen Variablen wird in der kanonischen Variablenanalyse dadurch Rechnung getragen, daß als zugrunde liegendes Proximitätsmaß die Mahalanobis-Distanz verwendet wird, und diese durch euklidische Distanzen repräsentiert beziehungsweise approximiert (wenn q < p) wird (GOWER & HAND, 1996). Obwohl die kanonische Variablenanalyse in dieser Arbeit ausschließlich deskriptiv eingesetzt wird, ist zu beachten, daß der Methodik die Annahme der Varianzhomogenität, das heißt der Gleichheit der Kovarianzmatrizen der einzelnen Gruppen, inhärent ist, da eigentlich nur dann die Bildung einer gepoolten SSP-Matrix **W** für die Streung innerhalb der Gruppen sinnvoll ist (KRZANOWSKI, 1988a).

---

[26] Daher auch die Bedeutung der kanonischen Variablen in der Diskriminanzanalyse, auf die in dieser Arbeit jedoch nicht eingegangen wird (siehe zum Beispiel KRZANOWSKI & MARRIOTT, 1994 und 1995)

2.3.3.3 Nichtlineare kanonische Analyse

Die nichtlineare, generalisierte, kanonische Analyse, die auf GIFI, 1990, zurückgeht, kann als Verallgemeinerung der kanonischen Korrelationsanalyse verstanden werden; das heißt, es geht um die Bestimmung der Beziehungen von Variablensets, also um den Versuch durch die gleichzeitige Betrachtung der Beziehungen der Variablensets untereinander, so viel wie möglich der vorhandenen Variabilität durch Linearkombinationen der Variablensets zu 'erklären'. Im Gegensatz zur linearen Korrelationsanalyse können aber mehr als zwei Variablensets gleichzeitig betrachtet werden und Variablen, die auf beliebigen Skalenniveaus vorliegen, können in die Analyse miteinbezogen werden, nicht nur intervall- und verhältnisskalierte Variablen wie in der linearen, kanonischen Korrelationsanalyse (HEISER & MEULMANN, 1995). Eine Umsetzung der Methodik in statistische Software liegt mit dem Programm OVERALS vor (SPSS, 1994).

Die generalisierte kanonische Analyse zählt zu den Verfahren der optimalen Skalierung, die im wesentlichen durch drei Aspekte charakterisiert sind (GIFI, 1990). Zum einen beinhalten diese Verfahren beliebige, nichtlineare Transformation der Ausgangsvariablen, die zuvor in eine Indikatormatrix (siehe 2.1.3) umgewandelt werden. Liegen die Variablen nicht ursprünglich nominal- oder ordinalskaliert vor ist also eine entsprechende Klassenbildung vorzunehmen, zum Beispiel eine einfache Rangtransformation. Das zweite Charakteristikum ist die Verwendung eines alternierenden, kleinste Quadrate Algorithmus zur Ermittlung der Objektwerte[27] und Variablenquantifikationen, das heißt der optimalen Transformation für die Ausgangsvariablen. Die Ermittlung der Variablenquantifikationen errfolgt auf iterativem Weg, wobei die Abweichungen zwischen Objektwerten und den Werten der quantifizierten Ausgangsvariablen in einer gewählten Dimensionalität minimiert werden. Schließlich, und das ist der dritte Aspekt, können in der Analyse bestimmte Begrenzungen im Rahmen der Transformationen vorgegeben werden, je nachdem auf welchem Skalenniveau die Daten analysiert werden sollen, unabhängig vom Skalenniveau, auf dem sie gemessen werden. Unterschieden wird zwischen numerischem und ordinalen, sowie einfach und mehrfach nominalen Skalenniveau (genaue Definition siehe SPSS, 1994, KRZANOWSKI & MARRIOTT, 1994). Zur Interpretation der Lösung der nichtlinearen kanonischen Analyse kann die multiple Anpassung (multiple fit) der Variablen berechnet werden, die angibt, wie stark die Dimensionen durch die einzelnen Variablen beeinflußt werden und welche Variablen den stärksten diskriminatorischen Beitrag zur Trennung der Objekte liefern. Ebenfalls informativ sind die Komponentenladungen, die gleich den Korrelationen der quantifizierten Ausgangsvariablen und der Objektwerte sind. Die Loss-Werte je Variablenset geben schließlich an, wie gut beziehungsweise wie schlecht die Übereinstimmung zwischen den im Algorithmus ermittelten Objektwerten und den Objektwerten bei Verwendung der optimal quantifizierten Variablen ist. Die Minimierung dieses

---

[27] Vergleichbar den Hauptkomponentenwerten für die Objekte in der Hauptkomponentenanalyse.

Unterschieds ist das Ziel des Iterationsprozesses der nichtlinearen kanonischen Analyse. Das Gegenstück zum Loss ist der Fit der Analyselösung. Loss und Fit summieren sich zur Anzahl der betrachteten Dimensionen. Die Loss-Werte sind, ähnlich wie die stress-Werte in der ordinalen mehrdimensionalen Skalierung, ein Maßstab für die Güte der dimensionserniedrigten Darstellung, je niedriger der Loss ist, desto besser ist die Darstellung.

2.4 Linienverbände

2.4.1 Formale Begriffsanalyse

Ein Begriff wird in der Philosophie als eine gedankliche Einheit mit einem bestimmtem Begriffsinhalt und einem bestimmten Begriffsumfang verstanden. In der formalen Begriffsanalyse geht es um die mathematische Formalisierung dieses Begriffsverständnisses mit der Bereitstellung eines flexiblen Instruments der Wissenskommunikation. WILLE, 1982, gilt als Begründer der formalen Begriffsanalyse, die von ihm und der Forschungsgruppe Begriffsanalyse der Technischen Hochschule Darmstadt kontinuierlich weiterentwickelt wird. Von ihren Entwicklern wird die formale Begriffsanalyse als eine in der pragmatischen Philosophie verwurzelte Methode verstanden, die ein besonderes Augenmerk darauf richtet, daß die entwickelte Methodik immer und nachvollziehbar im Bezug zur Wirklichkeit steht. Die formale Behandlung von Daten soll sich demnach nicht vom allgemeinen Verständnis der Daten lösen. Auf eine einfache Rekonstruktion der in der Analyse verwendeten Orginaldaten wird Wert gelegt, damit bei der Interpretation der Analyseergebnisse der ursprüngliche, inhaltliche Zusammenhang immer faßbar bleibt (KOLLEWE et al., 1994).

Zunächst sollen die konzeptionellen Grundlagen der formalen Begriffsanalyse kurz dargestellt werden (2.4.1.1). Die für die Datenanalyse wichtigen Liniendiagramme, die das wesenliche Kommunikationsintrument der formalen Begriffsanalyse sind, werden unter 2.4.1.2 erläutert. Umfangreiche Datensätze lassen sich zweckmäßiger mit gestuften als mit einfachen Liniendiagrammen nach erfolgter begrifflicher Skalierung visualisieren (siehe 2.4.1.3). Die Darstellung in den folgenden Kapiteln erfolgt in Anlehnung an den Sprachgebrauch der Forschungsgruppe Begriffsanalyse.

2.4.1.1 Konzeptionelle Grundlagen

Ein Begriff ist gekennzeichnet durch einen gewissen Begriffsumfang, das heißt durch alle Objekte oder Gegenstände, die zum Begriff gehören, beziehungsweise durch alle Merkmale oder Variablenausprägungen, die zum Begriff zählen. Werden alle Gegenstände g als Elemente einer Menge G und alle Merkmale m als Elemente einer Menge M bezeichnet, so ist ein formaler Kontext definiert durch K:=(G,M,I), wobei I für die binäre Relation zwischen den Elementen G und M steht und geschrieben werden kann gIm, sprich der Gegenstand g besitzt das Merkmal m. Ein formaler Begriff des formalen Kontext (G,M,I) mit dem Begriffsumfang A und dem Begriffsinhalt B, ist das Paar (A,B), für das gelten: $A \subseteq G$ (sprich: A Teilmenge von G) und $B \subseteq M$, sowie A = B' und B = A', wobei B' in dieser Definition die Menge der gemeinsamen Merkmale der Gegenstände A (des Begriffsumfangs) ist, und A' die Menge der Gegenstände, die alle die Merkmale B, das heißt den gleichen Begriffsinhalt, besitzen. Ein formaler Kontext ist also dadurch gekennzeichnet, daß man immer von der Menge der gemeinsamen Merkmale der Gegenstände eines Begriffes, zur Menge der Gegenstände, die diese Merkmale gemeinsam besitzen, gelangt.

Die Darstellung eines formalen Kontextes erfolgt als Kreuztabelle, wie zum Beispiel in Tabelle 4.

Tabelle 4: Beispiel einer Kreuztabelle eines formalen Kontext

	Topfkultur	Schnittkultur	Kultur < 1 Jahr	Kultur > 1 Jahr
Gerbera		x		x
Dianthus		x		x
Chrysanthemum		x	x	
Pelargonium	x		x	
Lilium		x	x	

Ist nun zum Beispiel A {Gerbera, Dianthus}, so ist A' {Schnittkultur, Kultur > 1 Jahr}. Für B {Schnittkultur, Kultur > 1 Jahr} gilt B' {Gerbera, Dianthus}. Der Begriff (A,B), in diesem Beispiel, hat also den Begriffsumfang Gerbera und Dianthus und den Begriffsinhalt Schnittkultur mit mehr als einem Jahr Kulturdauer.

Weiter gilt, daß es unter allen Begriffen eines Kontextes eine hierarchische Ordnung in Ober- und Unterbegriffe mit folgender Beziehung gibt: $(A_1, B_1) \subseteq (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq B_1)$, das heißt also, wenn der Begriff (A_1, B_1) ein Unterbegriff des Begriffes (A_2, B_2) ist, so folgt daraus, das die Gegenstandsmenge A_2 die Gegenstandsmenge A_1 einschließt, und die Merkmalsmenge B_1 die Merkmalsmenge B_2 einschließt. So ist im Beispiel der Begriff A_2 {Gerbera, Chrysanthemum, Dianthus}, B_2 {Schnittkultur}, ein Oberbegriff von A_1 {Chrysanthemum}, B_1 {Schnittkultur, Kultur < 1 Jahr}. Die Ordnung aller Begriffe eines formalen Kontext ergibt einen Begriffsverband, der durch ein beschriftetes Liniendiagramm darstellbar ist, das in der Regel durch spezielle Algorithmen am Computer, und nur in sehr kleinen Datensätzen mit der Hand erstellt wird (ESZ, 1996, WILLE, 1987, WOLFF, 1988 und 1993).

WILLE, 1987, gibt einige Hinweise zum möglichen Einsatz von Begriffsverbänden, so unter anderem die hierarchische Klassifikation von Gegenständen (Objekten), die Untersuchung von Merkmalsimplikationen, die Bereitstellung einer Struktur zur Darstellung und Abfrage von Wissen oder die Bestimmung von Gegenständen. Der Ansatz der Klassifikation und Gruppierung von Gegenständen ist für die Datenanalyse neben der strukturierten Bereitstellung von Wissen wohl der wichtigste Bereich. Beispiele für praktische Anwendungen bieten SPANGENBERG & WOLFF, 1991, WOLFF, 1993, und WOLFF & STELLWANGEN, 1992. SPANGENBERG & WOLFF, 1991 stellen dabei Biplots und formale Begriffsanalyse in psychologischen Untersuchungen einander

gegenüber. LENGNINK, 1993, gibt eine Darstellung zur Behandlung von Proximitätsmatrizen.

2.4.1.2 Einfache Liniendiagramme

Ein Liniendiagramm ist die graphische Realisation eines Begriffsverbandes. Ein Beispiel, aufbauend auf dem Kontext in Tabelle 4 zeigt Abbildung 1. Die Punkte des Liniendiagrammes stehen für die Begriffe des Kontextes, die Linien zwischen den Punkten verdeutlichen die hierachische Ordnung des Begriffsverbandes, das heißt eine aufsteigende Linie verbindet Unterbegriff mit Oberbegriff. Der oberste Punkt steht für einen alle Merkmale und Gegenstände umfassenden Begriff, der

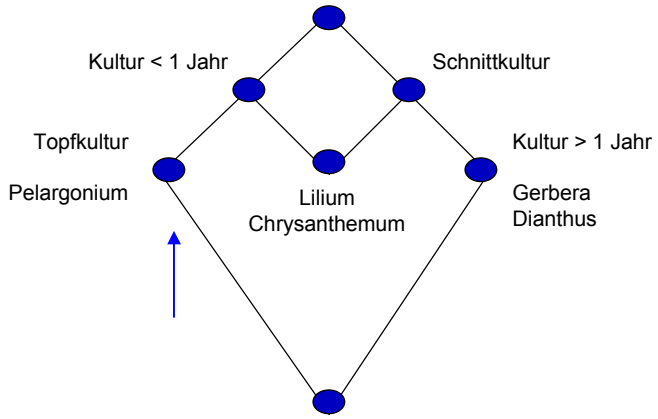


Abbildung 1: Einfaches Liniendiagramm, Daten aus Tabelle 4

unterste Punkt für einen (möglicherweise) weder Gegenstände noch Merkmale beinhaltenden Begriff.

Nicht jeder Begriff des Liniendiagramms muß beschriftet werden. Ein Punkt wird mit einem Gegenstandsnamen versehen, wenn dieser Kreis den Begriff γg repräsentiert, das heißt den Begriff mit dem kleinsten Begriffsumfang, der den Gegenstand g enthält. Ebenso wird ein Punkt mit einem Merkmalsnamen versehen, wenn dieser Punkt den Begriff μm repräsentiert, das heißt den Begriff mit dem größten Begriffsinhalt, der das Merkmal m enthält. Es ergibt sich so die Leseregel, daß der Umfang eines Begriffes durch alle Gegenstände definiert ist, die auf einer absteigenden Linie vom Punkt des Begriffes aus zu erreichen sind, und daß der Inhalt eines Begriffes durch alle Merkmale definiert ist, die auf einer aufsteigenden Linie vom Punkt des Begriffes aus erreicht werden können. Im Beispiel in Abbildung 1 ergibt sich also für den mit dem Pfeil gekennzeichneten Begriff der Begriffsumfang pelargonium und der Begriffsinhalt topfkultur und kultur < 1 jahr.

Als Folge der hierarchischen Ordnung ergibt sich zudem, daß ein Gegenstand genau alle diejenigen Merkmale besitzt, die mit einer aufsteigenden Linie vom Begriff, der die

Gegenstandsbezeichnung trägt, erreicht werden können, und daß ein Merkmal allen Gegenständen gemein ist, die mit einer absteigenden Linie vom Begriff, der die Merkmalsbezeichnung trägt, erreicht werden können. Das Liniendiagramm bildet somit den formalen Kontext ohne Informationsverlust ab. Je größer die Kontexte werden, desto schwerer lesbar wird jedoch das Liniendiagramm und es bietet sich die Verwendung gestufter Liniendiagramme an (WOLFF, 1993).

2.4.1.3 Begriffliches Skalieren und gestufte Liniendiagramme

In der Mehrzahl der auszuwertenden Daten handelt es sich nicht um einwertige, sondern um mehrwertige Kontexte, das heißt ein Merkmal kann zwei oder mehr Ausprägungen annehmen. Dies betrifft sowohl nominalskalierte und ordinalskalierte als auch, und in besonderem Umfang, verhältnis- oder intervallskalierte Variablen. Während jedoch bei nominal- und ordinalskalierten Variablen die einzelnen Merkmalsausprägungen bereits vorgegeben sind, ist bei der Bearbeitung von intervall- oder verhältnisskalierten Variablen eine gesonderte Klassenbildung vorzunehmen. Die Klassenbildung und die Bearbeitung des Kontexts im Sinne der Klassenbildung wird in der formalen Begriffsanalyse als begriffliche Skalierung bezeichnet. Die Auswahl einer geeigneten Skala richtet sich nach der Fragestellung in der jeweiligen Untersuchung. Die begriffliche Skalierung erfordert daher eine enge Zusammenarbeit zwischen dem sogenannten Präparator, der die technische und mathematische Aufarbeitung durchführt und dem eigentlichen Nutzer, der primär an den inhaltlichen Ergebnissen interessiert ist.

Der erste Schritt stellt die Entwicklung abstrakter Skalen dar. Verschiedene Grundtypen abstrakter Skalen sind in Tabelle 5 und Abbildung 2 kurz aufgeführt. Diese Auflistung ist natürlich nicht vollständig, und im Prinzip ist eine beliebig große Anzahl unterschiedlicher Skalentypen denkbar. Werden die abstrakten Skalen mit Bezeichnungen der Merkmale einer konkreten Datenbasis und den Deskriptoren[28] der Gegenstände versehen, so entstehen die konkreten Skalen (häufig führt natürlich auch der Weg von der konkreten zur abstrakten Skala). Die Zuordung von Gegenständen zu den Deskriptoren, entsprechend des mehrwertigen Kontexts, führt dann zu der realisierten Skala.

Die Gesamtheit der für den zu untersuchenden Kontext zutreffenden realisierten Skalen kann in Form gestufter Liniendiagramme dargestellt beziehungsweise nach und nach erkundet werden. Ein gestuftes Liniendiagramm entsteht, zum Beispiel im Fall von zwei Variablen, durch Ineinanderfügen mehrerer Begriffsverbände in der Art, daß das Liniendiagramm der einen Variablen ‘aufgeblasen’ wird (dies ergibt die Grobstruktur), und das Liniendiagramm der zweiten Variablen in dieses erste Liniendiagramm eingefügt wird (dies ergibt die Feinstruktur). Beispiele sind mehrfach in Kapitel 3 zu finden. Die Leseregel bleibt identisch zu der des einfachen Liniendiagramms, das heißt, ein Begriff

---

[23] Deskriptoren stehen für eine Anzahl von Gegenständen, auf die über die Deskriptoren zugegriffen werden kann.

b_1 ist ein Unterbegriff des Begriffes b_2, wenn b_1 sowohl in der Grobstruktur als auch in der Feinstruktur ein Unterbegriff von b_2 ist. Die gleichzeitige Darstellung von mehr als drei Variablen führt aber häufig schon, wie beim einfachen Liniendiagramm, zu einem nicht mehr lesbaren Bild. Um einen komplexen, mehrwertigen Kontext daher mit Hilfe der formalen Begriffsanalyse zu verstehen, ist ein interaktives, den Kontext nach und nach erkundendes Vorgehen erforderlich. Die notwendige Software liegt mit dem Programm TOSCANA vor (KOLLEWE et al., 1994, NAVICON, 1996).

Tabelle 5: Einige typische Skalen in der begrifflichen Skalierung

Ordinalskala

Merkmal/ Gegenstand	>=1	>= 2	>= 3	>= 4
1	X			
2	X	X		
3	X	X	X	
4	X	X	X	X

Biordinalskala

Merkmal/ Gegenstand	<= 3	<= 2	<= 1	>= 4	>= 5	>= 6
1	X	X	X			
2		X	X			
3			X			
4				X		
5				X	X	
6				X	X	X

Interordinal-skala

Merkmal/ Gegenstand	billig	nicht teuer	mittel	nicht billig	teuer
1	X	X			
2		X	X	X	
3				X	X

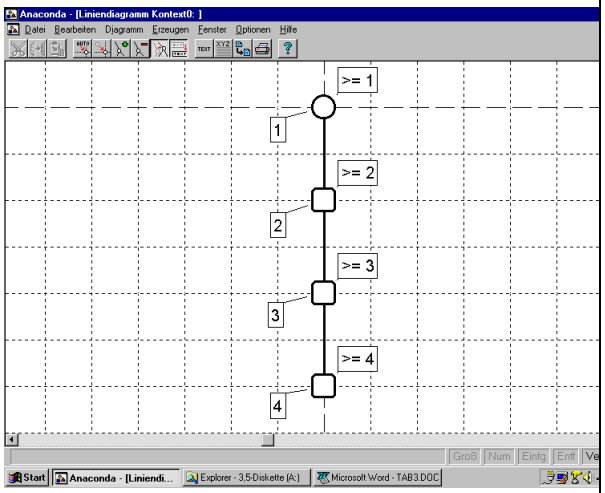
Dichotome Skala

Merkmal/ Gegenstand	männlich	weiblich
1	X	
2		X

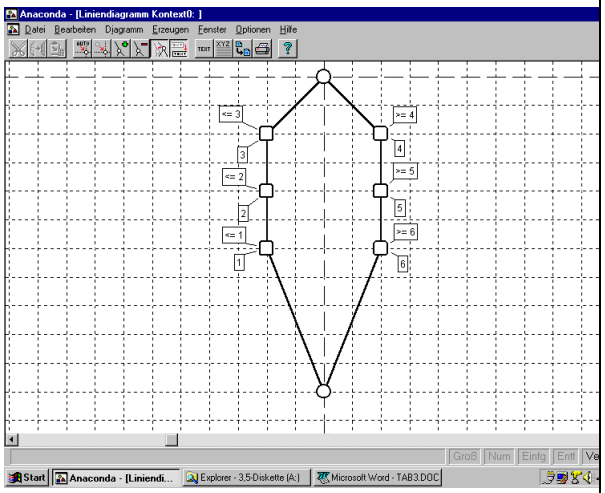
Nominalskala

Merkmal/ Gegenstand	blau	gelb	grün
1	X		
2		X	
3			X

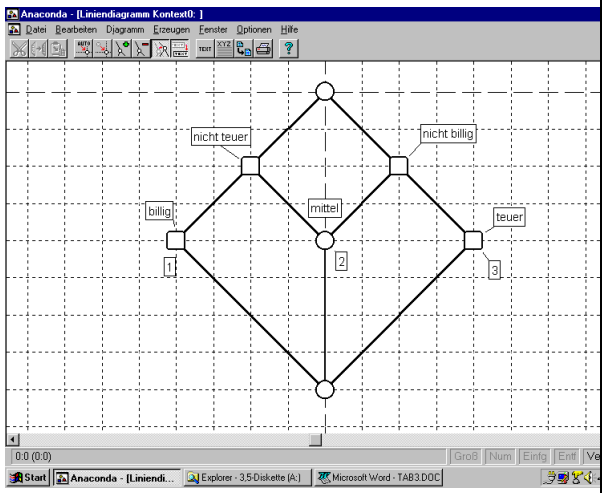
Ordinalskala



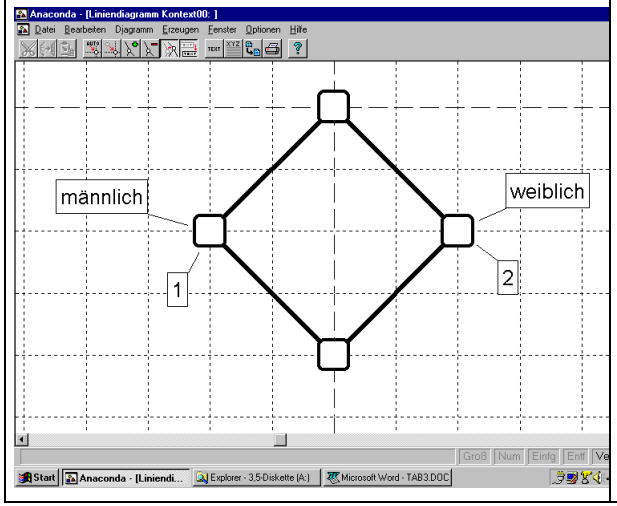
Biordinalskala



Interordinalskala



Dihotome Skala



Nominalskala

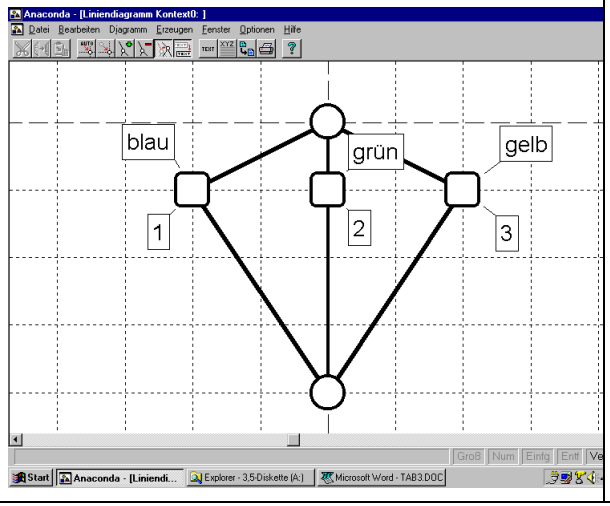


Abbildung 2: Einige typische Skalen in der begrifflichen Skalierung

2.4.2 Graphische Modelle

Graphische Modelle dienen der Untersuchung und Darstellung multivariater Beziehungszusamenhänge auf Grundlage der bedingten Unabhängigkeit. Bedingte Unabhängigkeit ist zum Beispiel für die Variablen A, B und C gegeben, wenn gilt: $A \perp\!\!\!\perp B|C$, sprich A unabhängig B, gegeben C. Das Konzept der bedingten Unabhängigkeit überwindet die Probleme, die bei der paarweisen Betrachtung von Variablen auftreten können und die als Paradoxum nach Simpson bekannt sind (SIMPSON, 1951). Die Standardliteratur zu graphischen Modellen gibt verschiedene Beispiele für vorgetäuschte Beziehungen, die sich durch die zusätzliche Betrachtung einer weiteren Variablen als solche herausstellen (siehe zum Beispiel EDWARDS, 1995 oder WHITTAKER, 1990).

Zur Darstellung der Ergebnisse des graphischen Modellbildungsprozeses werden gerichtete oder ungerichtete Graphen oder Graphen mit gerichteten und ungerichteten Verbindungen verwendet, die ihre Quellen in der Graphentheorie haben (LAURITZEN, 1996)[29]. Zwei Variablen in einem graphischen Modell sind bedingt unabhängig, wenn sie nicht durch eine direkte Linie miteinander verbunden sind. Beispiele sind im Auswertungsteil zu finden.

Im Gegensatz zu der Mehrzahl der in dieser Arbeit besprochenen und eingesetzten Methoden, handelt es sich bei graphischen Modellen um im statistischen Sinne echte Modelle, das heißt, es werden Modelle gebildet, die die Beziehungen zwischen den untersuchten Variablen repräsentieren und deren Angemessenheit mit Hilfe probabilistischer Verfahren überprüft wird. Insofern sind graphische Modelle nicht frei von Annahmen, zum Beispiel zur Verteilung der Daten[30]. Vielmehr basiert der Modellbildungsprozeß auf der Durchführung von Signifikanztests zur Auswahl des oder der adäquaten Modelle (siehe unten), wobei die Richtigkeit oder Angemessenheit eines Modells natürlich auch und vor allem unter sachlogischen Gesichtspunkten zu betrachten ist und es das eine und richtige Modell für die zu untersuchenden Daten nicht geben kann. Diese Unsicherheit im Modellbildungsprozeß wird vor allem durch den EH-Algorithmus verdeutlicht (EDWARDS, 1995, siehe unten).

Je nach Datenherkunft lassen sich diskrete, kontinuierliche und gemischte graphische Modelle einsetzen. Diskrete graphische Modelle untersuchen die Wahrscheinlichkeiten der Zellhäufigkeiten von 2-, 3- oder Mehr-Wegetafeln nominal- oder ordinalskalierter Variablen. Sie sind eine

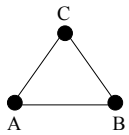
---

[29] In dieser Arbeit werden, bedingt durch die zu untersuchenden Daten (siehe Kapitel 3), ausschließlich ungerichtete Graphen eingesetzt. Eine ausführliche Behandlung gerichteter Graphen geben COX & WERMUTH, 1996.

[30] Über die geringe Qualität der in dieser Arbeit untersuchten Daten wird bereits in der Einführung hingewiesen. Die verrechneten Daten stellen in keinem Fall eine repräsentative Stichprobe einer hypothetischen Grundgesamtheit dar. Insofern sind die graphischen Modelle im Auswertungsteil auch ausschließlich deskriptiv und explorativ zu verstehen.

Unterordnung aller möglichen log-linearen Modelle (FIENBERG, 1980), deren Besonderheit darin liegt, daß, wenn zwischen zwei Variablen Unabhängigkeit festgestellt wird, also, um in der Sprache der log-linearen Modelle zu bleiben, die Zwei-Faktor-Wechselwirkung auf Null gesetzt wird, alle höherwertigen Wechselwirkungen, die diese Variablen beinhalten, ebenfalls gleich Null gesetzt werden. Höherwertige Wechselwirkungen werden also durch die Zwei-Faktor-Wechselwirkungen bestimmt. Wenn zum Beispiel im Fall der oben genannten drei Variablen A, B und C gilt, daß die Wechselwirkung zwischen B und C (BC) nicht signifikant ist, also gleich Null gesetzt wird, so gilt automatisch, daß die Drei-Faktor-Wechselwirkung ABC auch gleich Null gesetzt wird.

Es gibt hierarchische log-lineare Modelle, die nicht graphisch sind. Das log-lineare (gesättigte) Modell ABC mit dem Graphen



ist graphisch, es beinhaltet alle Zwei-Faktor-Wechselwirkungen und damit auch die Drei-Faktor-Wechselwirkung. Das log-lineare Modell AB, AC, BC, ohne Drei-Faktor-Wechselwirkung, aber mit gleichem Graph, ist demgegenüber ein nicht graphisches, hierarchisches log-lineares Modell, da die Drei-Faktor-Wechselwirkung fehlt, obwohl alle Zwei-Faktor-Wechselwirkungen vorhanden sind (EDWARDS, 1995).

Kontinuierliche graphische Modelle dienen zur Analyse multinormalverteilter intervall- oder verhältnisskalierter Variablen, das heißt sie setzen das Vorliegen der Multinormalverteilung voraus. Zwei Variablen in kontinuierlichen graphische Modellen sind voneinander bedingt unabhängig, wenn die partiellen Korrelationen zwischen diesen Variablen, gegeben die übrigen Variablen, nicht signifikant sind, oder, was das gleiche ist, wenn die zu dem Variablenpaar gehörenden Eintragungen in der Inversen der Kovarianzmatrix (in der sogenannten Präzisionsmatrix) gleich Null gesetzt werden können. Im Gegensatz zu den diskreten graphischen Modellen gibt es keine hierarchischen, nicht-graphischen Modelle.

Der Einsatz gemischter graphischer Modelle ergibt sich bei der gleichzeitigen Verrechnung von diskreten und kontinuierlichen Daten. Die angenommene Verteilung der Daten entspricht der CG-Verteilung (Conditional Gaussian); das heißt, es wird angenommen, daß die Wahrscheinlichkeit, daß die diskrete Zufallsvariable *I* den Wert *i* annimmt (*I* = *i*), p_i ist, und daß die Verteilung der kontinuierlichen Zufallsvariablen Y, gegeben *I* = *i*, multivariat normal ist, mit Mittelwert μ_i und Kovarianzmatrix Σ_i, das heißt sowohl der Mittelwert als auch die Kovarianzmatrix sind bedingt durch *i*.

Um zu einem graphischen Modell zu gelangen, ist ein Modellbildungsprozeß notwendig, der sowohl

durch seine Vorgehensweise als auch durch die Auswahl eines bestimmten Hypothesentests charakterisiert ist. An Vorgehensweisen lassen sich die Rückwärts-Elimination, die Vorwärts-Selektion und der EH-Algorithmus unterscheiden. Die Rückwärts-Elimination geht vom vollen Modell aus (das heißt es bestehen Wechselwirkungen zwischen allen Variablen und damit direkte Verbindungen im Graphen) und entfernt sukzessive die am wenigsten signifikanten Verbindungen zwischen zwei Variablen. Der Anpassungsverlust beim Vergleich zweier aufeinanderfolgender, hierarchischer Modelle ist dann ein Maßstab für die Annahme oder Ablehnung des gebildeten Modells. Die Vorwärts-Selektion geht entsprechend vor, wählt jedoch als Ausgangspunkt das Modell völliger Unabhängigkeit zwischen den Variablen und fügt diesem Modell nach und nach die am höchsten signifikanten Variablenverbindungen zu, bis ein weiteres Hinzufügen keine signifikante Verbesserung des Modells mehr erbringt. Es ist offensichtlich, daß mit beiden Methoden zwar Modelle gefunden werden können, die zu einer mit den Daten vereinbaren Darstellung führen, daß aber auch eine Vielzahl an anderen Modellen, die ebensogut an die vorliegenden Daten angepaßt werden könnten, durch das schrittweise Vorgehen übersehen werden können. Eine Alternative bietet der EH-Algorithmus. Es handelt sich um einen Suchalgorithmus, der eine große Anzahl an Modellen untersucht und daraufhin testet, ob die Modelle mit den Daten vereinbar sind oder nicht, und die Modelle dann als mögliche Modelle akzeptiert oder zurückweist (EDWARDS & HAVRÁNEK, 1985 und 1987). Eine Diskussion über die Vor- und Nachteile der unterschiedlichen Selektionsverfahren gibt SMITH, 1992.

Schließlich ist eine Teststatistik für den Modellfindungsprozeß zu definieren. Verwendet wird in dieser Arbeit der G^2-Test. G^2 wird berechnet als Differenz zwischen zwei miteinander zu vergleichenden, hierarchischen diskreten graphischen Modellen durch:

$G^2 = 2\sum_{jkl} n_{jkl} \ln\left(\hat{m}^1_{jkl} / \hat{m}^0_{jkl}\right)$, wobei n_{jkl} die beobachtete Zellhäufigkeit in einer 3-Wege Tafel mit den diskreten Variablen A, B und C ist, die in die Klassen (j = 1 ... a), (k = 1 ... b) und (l = 1 ... c) eingeteilt sind, und $\hat{m}^1_{jkl}$ die Maximum Likelihood-Schätzung der Zellhäufigkeit unter Modell 1 (des einfacheren Modells) und $\hat{m}^0_{jkl}$ die Maximum Likelihood-Schätzung der Zellhäufigkeit unter Modell 0 (des komplexeren Modells) darstellt. G^2 folgt asymptotisch der Chi-Quadratverteilung mit k Freiheitsgraden, wobei k gleich der Differenz an Freiheitgraden von Modell 0 minus Anzahl an Freiheitsgraden von Modell 1 ist (zu den exakten n Definitionen und Alternativen zu G^2 siehe EDWARDS, 1995). Im Kontext dieser Arbeit ist vor allem zusätzlich darauf hinzuweisen, daß auf Grund der Vielzahl an Variablen und der im Vergleich zu den möglichen Variablenkombinationen geringen Zahl an Objekten, vielfach schwach besetzte Tabellen in diskreten graphischen Modellen mit vielen Zellen mit Nulleinträgen vorkommen. In einem solchen Fall ist auf exakte Testverfahren, zum Beispiel basierend auf Monte Carlo Simulationen, zurückzugreifen (Näheres zu exakten Tests in graphischen Modellen zum Beispiel in WHITTAKER, 1990).

Abschließend ist anzumerken, daß die allgemeinen, dem Verfasser zur Verfügung stehenden

Statistikprogramme (Genstat, S-Plus und SPSS), keine zufriedenstellende Behandlung graphischer Modelle ermöglichen. Eine gute Lösung bietet das Programm MIM, das für diese Arbeit nicht verfügbar ist. Da zudem ausschließlich diskrete graphische Modelle eingesetzt werden, erfolgt die Auswertung ausschließlich mit dem Programm DIGRAM (KREINER, 1989).

2.4.3 Regressions- und Klassifikationsbäume

Klassifikations- und Regressionsbäume (Baumdiagramme) bieten die Möglichkeit eine Menge von Objekten in möglichst homogene Segmente (Gruppen) zu unterteilen. Insofern besteht eine gewisse Ähnlichkeit zur Clusteranalyse (siehe 2.5.1.2). Baumdiagramme leisten aber, vor allem unter dem Gesichtspunkt der Datenvisualisierung, noch mehr. Die durch die Analyse entstehende Baumstruktur zeigt nämlich nicht nur auf, welche Segmente gebildet werden und welche Objekte den jeweiligen Segmenten zugeordnet werden, sondern auch, welche Variablen diese Segmente in erster Linie charakterisieren und welche Variablen aus der Anzahl aller, in einer Analyse betrachteten, Merkmalen, den stärksten segmentierenden Einfluß haben.

Klassifikations- und Regressionsbäume, die von BREIMAN et al., 1984, beschrieben werden und auch unter der Bezeichnung CART bekannt sind, eignen sich für die Analyse gemischter Datensätze, die sowohl diskrete (nominal- und ordinalskalierte) als auch kontinuierliche (intervall- und verhältnisskalierte) Variablen beinhalten. Um zu einem Baumdiagramm zu gelangen ist es zunächst erforderlich, eine Variable als die Zielvariable zu kennzeichnen. Ist die Zielvariable diskret, so wird von einem Klassifikationsbaum, ist sie kontinuierlich, von einem Regressionsbaum gesprochen. Eine diskrete Zielvariable sollte annähernd multinomialverteilt, eine kontinuierliche Zielvariable annähernd normalverteilt sein. Der Zielvariable gegenüber stehen die Prediktorvariablen, die ein beliebiges Skalenniveau aufweisen können und über die keine Verteilungsannahmen gemacht werden.

Das Verfahren, das zum Aufbau eines Baumdiagramms führt, wird als rekursive Partitionierung bezeichnet. Im ersten Schritt wird die Prediktorvariable gesucht, die bei einer Trennung der Objekte in zwei Gruppen zu einer möglichst großen Homogenität innerhalb und möglichst großen Heterogenität zwischen den Gruppen bezüglich der gewählten Zielvariablen führt. Diese Homogenität kann nach BREIMAN et al., 1984, zum Beispiel mit einem sogenannten Unreinheitsindex bestimmt werden; darüber hinaus existieren verschiedene andere Indizes, um den optimalen Aufspaltungswert zu bestimmen, die aber häufig zu sehr ähnlichen Ergebnissen führen. Die so gebildeten Segmente, die jetzt an einem sogenannten Terminalknoten liegen, werden ihrerseits nun wieder nach demselben Prinzip durch binäre Splits in zwei Untergruppen unterteilt, wobei im Laufe der Bildung des Baumdiagramms ein und dieselben Variablen an verschiedenen Stellen auftauchen können. Der Baum kann solange weiterwachsen, bis an einem Terminalknoten nur noch ein Objekt beziehungsweise nur Objekte mit identischen Werten bei der Zielvariablen vorliegen, so daß eine weitere Aufsplittung nicht möglich ist. Häufig stoppt der Entwicklungsprozeß jedoch schon früher, und zwar wenn eine bestimmte Anzahl Objekte an einem Terminalknoten unterschritten wird. Nach NAGEL et al., 1996, empfiehlt es sich keine weiteren Splits an einem Terminalknoten vorzunehmen, wenn bei n Objekten die Anzahl Objekte an einem Terminalknoten $< \sqrt{n}$ ist. Der Schätzwert der Zielvariablen in einem Regressionsbaum errechnet sich als der Mittelwert der Zielvariablen der Objekte im Segment. Die Residuen sind die quadrierten Differenzen

von Schätzwert und den beobachteten Werten der Objekte. Die Summe der Residuen aller Terminalknoten geteilt durch die Anzahl der Terminalknoten, wird als mittlere Residuendevianz (mean residual deviance) bezeichnet.

Um das Baumdiagramm übersichtlicher zu gestalten, ist es angebracht, den Baum zu ‘schneiden’, das heißt untere Terminalknoten bis zu einem gewissen Punkt zu entfernen, so daß Segmente an den Terminalknoten entstehen, die noch weiter unterteilt werden könnten, darauf aber verzichtet wird, um die wesentlichen Aspekte des Baumdiagramms stärker hervorzuheben. An welcher Stelle jedoch ein Baumdiagramm optimal ‘geschnitten’ ist, kann nicht eindeutig beantwortet werden. Ein Hilfsmittel, sich einer sinnvollen Baumgröße zu nähern, ist das sogenannte cost complexity pruning. Je mehr Terminalknoten betrachtet werden, desto geringer ist die mittlere Residuendevianz. Der Grundgedanke des cost complexity pruning ist es nun, eine Abfolge von Baumstrukturen zu finden, die bei einer gegebenen Anzahl an Terminalknoten (in der Regel von maximal bis minimal möglicher Anzahl an Terminalknoten), die jeweilige Struktur mit der geringsten mittleren Residuendevianz sind. Mit Hilfe einer Graphik der mittleren Residuendevianzen auf der y- und der Anzahl an Terminalknoten auf der x-Achse läßt sich dann abschätzen, an welcher Stelle es zu starken Sprüngen, das heißt starken Zunahmen in der mittleren Residuendevianz kommt und ansatzweise entscheiden, ob der Zugewinn an Einfachheit der Darstellung das weitere Anwachsen der mittleren Residuendevianzen noch wert ist (MATHSOFT, 1997). BREIMAN et al., 1984, geben weitere, auch numerische Hilfsmittel für die Auswahl des geeigneten Baumdiagramms.

Der CHAID (Chi Square Automatic Interaction Detector)-Algorithmus kann als Spezialfall der Klassifikations- und Regressionsbäume angesehen werden . Wesentliche Unterschiede lassen sich wie folgt zusammen fassen (KASS, 1980, SPSS, 1993):

1. CHAID verwendet ausschließlich diskrete Variablen, sowohl als Ziel- als auch als Prediktorvariablen;
2. an Stelle von binären Splits können Splits in eine beliebige Anzahl von Klassen vorgenommen werden.
3. die Splits orientieren sich an der Wahrscheinlichkeit aller Zwei Wege-Tafeln der Ziel- und der Prediktorvariablen; im ersten Schritt wird die Prediktorvariable ausgewählt, die die stärkste Assoziation zur Zielvariablen aufweist. In den dann entstandenen Segmenten wird dieses Vorgehen bis zum Ende der Entwicklung des Baumdiagramms wiederholt;
4. bei Bedarf können zwei Klassen einer Variablen in eine Klasse zusammengelegt werden, wenn deren Beziehungen zur Zielvariablen annähernd gleich sind; bei nominalen Variablen kann in der Regel eine freie Kombinierbarkeit der Klassen angenommen werden, während die Klassen ordinaler Variablen in der Regel monoton, das heißt nur mit einer direkt benachbarten Klasse in eine Klasse zusammengelegt werden dürfen;

5. die Entscheidungen zur Zusammenlegung von Klassen oder zur Auswahl der Prediktorvariablen an den Terminalknoten erfolgt auf der Grundlage von Chi-Quadrat-Tests der beobachteten und geschätzten Zellhäufigkeiten von Zwei-Wege-Tafeln mit der Ziel- und den Prediktorvariablen;
6. liegt eine ordinalskalierte Zielvariable vor, kann den ordinalen Klassen ein Wert zugeordnet werden. Sind die ordinalen Klassen zum Beispiel durch Transformation aus einer kontinuierlichen Variablen entstanden, kann jeder Klasse ihr Mittelwert oder Median zugeordnet werden. Sie werden in den Baumdiagrammen ausgewiesen und gehen auch in die Berechnung der Teststatistiken mit ein;
7. durch die sogenannte Bonferroni-Anpassung in den Chi-Quadrat Tests wird der Tatsache Rechnung zu tragen, daß es sich bei den Tests im Laufe der Entwicklung des Klassifikationsbaums nicht um voneinander unabhängige Tests handelt; sie bewirkt eine Verringerung des Signifikanzniveaus im einzelnen Test, um den nominalen Fehler aller Tests am festgelegten Signifikanzniveau zu halten.

2.5 Graphische und ergänzende Verfahren

2.5.1 Graphische Verfahren

In der Folge werden einige überwiegend graphisch eingesetzte Techniken angesprochen, die die bislang besprochenen Methoden ergänzen. Da sie an verschiedenen Stellen der Datenanalyse in Kapitel 3 eingesetzt werden, ist eine kurze Erwähnung angebracht; eine ausführliche Diskussion erfolgt jedoch nicht. Es werden besprochen

1. Andrews-Kurven und Parallelkoordinatenplots (2.5.1.1)
2. Dendrogramme und Multiple Spanning Trees (2.5.1.2)
3. Scatterplots und Trellis-Displays (2.5.1.3)
4. Interaktive Graphik und sonstige Verfahren (2.5.1.4)

2.5.1.1 Andrews-Kurven und Parallelkoordinatenplots

a-Andrews-Kurven

Andrews-Kurven gehen zurück auf ANDREWS, 1972. Jedem Objekt entspricht eine Andrews-Kurve, die als Funktion von $\omega(-\pi \le \omega \le \pi)$ nach dem folgenden Prinzip berechnet wird: $f_{x_{ij}}(\omega) = x_{i1} / \sqrt{2} + x_{i2} \sin\omega + x_{i3} \cos\omega + x_{i4} 2\sin\omega + x_{i5} \cos 2\omega + \ldots$, wobei die Anzahl der Variablen durch die Ordnung des Polynoms $f_{x_{ij}}(\omega)$ bestimmt wird. Ein Plot mit den Kurven jedes Objekts im Bereich von $-\pi$ bis π ergibt den Andrews-Plot. Wichtige Eigenschaften der Andrews-Kurven sind:

- die Funktionsrepräsentation der Objekte erhält den Mittelwert;
- die Funktionsrepräsentation der Objekte erhält die Varianz;
- die Funktionsrepräsentation der Objekte erhält die euklidische Distanz,

wobei gilt $D_{rt}^2 = \pi d_{rt}^2$, mit D_{rt}^2 als der quadrierten euklidischen Distanz zwischen zwei Funktionen und d_{rt}^2 als der quadrierten euklidischen Distanz zwischen zwei Objekten[31]. Das heißt also, daß zwei Kurven, die nahe beieinander liegen, auch im Sinne der euklidischen Distanz nahe beieinander sind. So können Andrews-Kurven helfen, Gruppierungen oder sehr aus dem allgemeinen Rahmen fallende Objekte aufzuspüren. Hilfreich ist bisweilen auch der Andrews-Plot an einem bestimmten Punkt ω. Als begrenzender Faktor für den Einsatz von Andrews-Kurven ist die Tatsache anzusehen, daß schon bei einer nur moderaten Anzahl an Objekten ein recht undeutliches Bild entstehen kann. Ein zweites Problem liegt darin begründet, daß die Reihenfolge, in der die Variablen in die Funktion eingehen, Einfluß auf den Funktionswert hat. Die ersten

---

[31] und zwar $D_{rt}^2 \int_{-\pi}^{\pi} (f_{x_{rj}}(\omega) - f_{x_{tj}}(\omega))\omega\omega$, wobei die Indices r und t für zwei Objekte r und t stehen.

Variablen haben in der Darstellung eine stärkeres Gewicht als später folgende Variablen, so daß es ratsam ist die Reihenfolge der Variablen so zu gestalten, daß die wichtigsten Variablen am Anfang stehen. Wo eine natürliche Reihenfolge nicht gegeben ist, ist die Durchführung einer Hauptkomponentenanalyse und die Bildung der Andrews-Kurven auf Grundlage der Hauptkomponentenwerte in Erwägung zu ziehen. Ein Beispiel für dieses Vorgehen liefert zum Beispiel ROVAN, 1994.

b-Parallelkoordinatenplots

In Parallelkoordinatenplots (WEGMAN, 1990) werden die Variablen durch parallele, vertikale oder horizontale Achsen dargestellt. Die Werte, die ein Objekt bei den jeweiligen Variablen einnimmt, werden durch eine Linie miteinander verbunden. So ist es möglich die Informationen zu einer Vielzahl von Variablen und Objekte ohne Informationsverlust in einer Abbildung unterzubringen. Je nach Variablenstruktur können die Originalwerte oder transformierte Werte beziehungsweise die absoluten oder die prozentualen Werte verwendet werden.

Parallelkoordinatenplots ermöglichen einen Einblick in die Korrelation der Variablen untereinander. Kommt es zum Überkreuzen der Objektlinien, so spricht dies für eine negative Korrelation; liegt ein paralleler Verlauf vor, so läßt dies den Schluß auf positive Korrelation zu. Da allerdings nahe beieinander liegende Parallelkoordinatenachsen (der Variablen) leichter Aufschluß über Korrelationen geben als weiter entfernt liegende, empfiehlt sich die Permutation der Variablenachsen. Um einen guten Überblick über mögliche Korrelationen zu erhalten, ist es allerdings nach KARAMAN, 1995, nicht erforderlich alle p! Permutationen der Variablenpaare von p Variablen abzubilden. Wenn erreicht wird, daß jede Variablenachse mindestens einmal neben jeder anderen Variablenachse plaziert wird, ist dies in der Regel ausreichend und bereits mit ungefähr p/2 Abbildungen zu erreichen.

Möglicherweise bei den Objekten vorhandene Gruppierungen lassen sich, ähnlich wie bei Andrews-Kurven, durch vergleichbare Linienverläufe unterschiedlicher Objekte identifizieren. Wie beim Andrews-Plot führt aber die Unübersichtlichkeit der Abbildungen bei vielen Objekten und die Vielzahl der Permutationen der Variablen zu einer begrenzten Nutzbarkeit der Parallelkoordinatenplots, sofern nicht interaktive Explorationsmöglichkeiten, wie sie zum Beispiel INSELBERG, 1997, beschreibt, eingesetzt werden können.

2.5.1.2 Dendrogramme und Multiple Spanning Trees

a-Dendrogramme und Clusteranalyse

Dendrogramme verdeutlichen graphisch die Ergebnisse einer hierarchischen Clusteranalyse[32]. Je

[32] Die Clusteranalyse stellt ein sehr umfangreiches Gebiet dar, das zum Beispiel von BACHER, 1994, ausführlich bearbeitet wird. Vorrangiges Ziel der Clusteranalyse ist die Gruppierung von

nach Clusterverfahren ergeben sich unterschiedliche Dendrogrammstrukturen, die sowohl Informationen zur Nähe beziehungsweise Entfernung von Objekten zueinander geben, als auch Aufschlüsse über mögliche Gruppierungen zulassen. Ein Dendrogramm ordnet die Objekte so an, daß einander ähnliche Objekte nahe beieinander, einander weniger ähnliche Objekte weiter von einander entfernt auf einer Linie liegen.

Es ist hierarchisch aufgebaut, das heißt Objektgruppierungen größerer Unähnlichkeit schließen Objektgruppierungen geringerer Unähnlichkeit ein. Geht die Gruppierung von einer Gesamtgruppe aus, die alle Objekte umfaßt und die nach und nach in Untergruppen unterteilt wird, so wird von einem divisiven Clusterverfahren gesprochen; geht die Gruppierung von einer der Anzahl der Objekte entsprechenden Zahl von Einzelgruppen (jedes Objekt entspricht also einer Gruppe) aus, die nach und nach durch weitere Objekte ergänzt wird, so liegt ein agglomeratives Clusterverfahren vor. Agglomerative Verfahren beherrschen die gängigen Vorgehensweisen in der Clusteranalyse, da sie weniger rechenintensiv als die divisiven Verfahren sind.

Ausgangspunkt für die Erstellung eines Dendrogramms ist eine, auf einem entsprechenden Proximitätsmaß beruhende, Proximitätsmatrix. Beim agglomerativen Vorgehen werden im ersten Schritt die beiden Objekte mit der geringsten Unähnlichkeit zu einer Gruppe zusammengefaßt; anschließend wird eine neue Proximitätsmatrix mit der neuen Gruppe an Stelle der zusammengefaßten Objekte berechnet und erneut auf der Grundlage dieser Proximitätsmatrix eine Zusammenführung von Objekten durchgeführt. Diese Schritte werden so lange wiederholt, bis nur noch eine Gruppe, die alle Objekte beinhaltet, vorliegt. Unterschiede zwischen hierarchischen Clusterverfahren beruhen nun auf unterschiedlichen Wegen, wie die Neuberechnung der Proximitätsmatrix nach der Zusammenführung von Objekten (wobei hier nun ein Objekt auch eine Gruppe von Objekten meinen kann) erfolgt. Einige Agglomerationskriterien sind in Tabelle 6 zusammengefaßt.

Weitere clusteranalytische Ansätze sind verschiedene Verfahren der modellbegründeten Clusteranalyse (BANFIELD & RAFTERY, 1992), sowie die nicht-hierarchische Klassifikation (Partitionierung um Medoide) und Fuzzy Clustering (KAUFMANN & ROUSSEEUW, 1990).

In der modellbegründeten Clusteranalyse wird mit Hilfe einer Maximum Likelihood Prozedur die Zuordnung eines Objekts zu einem Cluster (bei vorgegebener Clusterzahl) so vorgenommen, daß ein spezielles Kriterium optimiert wird, wobei das bekannteste wohl das Kriterium nach Ward ist, das zu einer Minimierung der Varianz innerhalb der gewählten Cluster führt. Andere Kriterien sind MATHSOFT, 1997, zu entnehmen. Die Anwendung unterschiedlicher Kriterien setzt unterschiedliche Annahmen zur Verteilung der Daten voraus (das Ward-Verfahren zum Beispiel die

---

Objekten aufgrund gemesener und beobachteter Merkmale. Methodische Einzelheiten werden in dieser Arbeit nicht besprochen.

Multinormalverteilung) und führt zu optimalen Ergebnissen unter der Annahme bestimmter Orientierungs-, Größen- und Formmerkmale der Cluster.

In der Partition um Medoide erfolgt die Clusterung, bei Vorgabe der gewählten Clusteranzahl, um spezielle, in den verschiedenen Clustern 'zentral angeordnete', repräsentative Objekte, den Medoiden. Diesen Medoiden werden weitere Objekte zugeordnet, die ihnen am ähnlichsten sind. Der Vorgang wird solange wiederholt, bis alle Objekte einem Cluster zugeordnet sind, und ein Austausch von Objekten zwischen unterschiedlichen Clustern zu keiner Verringerung der Summe der Unähnlichkeiten aller Objekte eines Clusters zum zugehörenden Medoid führt. K-means Clusterung geht entsprechend vor, verwendet aber statt einer Proximitätsmatrix die Originaldatenmatrix und minimiert nicht die Summe der Unähnlichkeiten, sondern die Summe der quadrierten, euklidischen Distanzen. Nach MATHSOFT, 1997, ist sie daher weniger robust als die Partition um Medoide.

Bei der Fuzzy Clusterung schließlich handelt es sich um eine unscharfe Gruppenzuordnung, das heißt die Objekte werden einem Cluster nur mit einer gewissen Wahrscheinlichkeit zugeordnet.

Da unterschiedliche Clusterverfahren zu unterschiedlichen Gruppierungen führen, unterliegen die Ergebnisse einer gewissen Beliebigkeit. Es gibt keine eindeutige Regel für das im Einzelfall geeignete und richtige Verfahren. Es ist zu beachten, daß die Clusteranalyse immer Objekte zu Gruppierungen zusammenfaßt, auch wenn den Objekten in Wirklichkeit überhaupt keine Gruppenstruktur zugrunde liegt. Jede Clusteranaylse teilt also eine (strukturierte oder unstrukturierte) Population in Gruppen ein. Zwei Fragen, die es daher vor Durchführung einer Clusteranalyse zu beantworten gilt, sind: „Kann überhaupt von einer Clusterung der Population ausgegangen werden?", und wenn ja: „Wieviel Cluster beschreiben die Population am besten?".

Eine Möglichkeit in der modellbegründeten Clusteranalyse die Anzahl der vorhandenen Cluster zu bestimmen und zu entscheiden, ob überhaupt eine Clusterstruktur vorliegt oder nicht, bietet die Berechnung sogenannter AWE[33]-Werte für jede Anzahl an möglichen Clustern (also von 1 bis n, mit n als der Anzahl der Objekte). Der höchste positive AWE-Wert gilt als Indiz für die Anzahl der in der Population tatsächlich vorhandenen Cluster. Liegen alle AWE-Werte unter Null, so ist dies ein Indiz, daß keine Clusterstruktur vorliegt.

Im Bereich der nicht-hierarchischen Klassifikation und der Fuzzy Clusterung kann die Erstellung von Silhouettenplots für eine unterschiedliche Anzahl von Clustern vorgenommen werden. Die Silhouettenbreite s(i) eines Objekts errechnet sich nach: $s(i) = b(i) - a(i) / \max\left[a(i), b(i)\right]$, mit a(i) als der mittleren Unähnlichkeit von Objekt i zu dem Cluster, dem es zugeordnet ist. Um b(i) zu berechnen ist es zunächst erforderlich die durchschnittliche Unähnlichkeit von Objekt i zu allen

[33] AWE = Approximate Weight of Evidence

übrigen gebildeten Clustern zu bilden. b(i) ist dann das Minimum dieser Unähnlichkeiten. Ein Wert von

s(i) = 1 entspricht einer sehr guten, ein Wert von s(i) = -1 einer sehr schlechten Klassifikation und der Wert s(i) = 0, deutet auf eine Lage des Objekts zwischen zwei Clustern hin. Im Silhouettenplot werden die Objekte nach ihren s(i) Werten sortiert wiedergegeben. Die mittlere Silhouettenbreite aller Objekte ist ein Hinweis auf die Güte Clusterlösung. Liegt sie unter 0,25, so ist dies ein Anzeichen für das Fehlen eine deutlichen Clusterstruktur. In der Fuzzy Clusterung kann zusätzlich der Dunn-Koeffizient betrachtet werden, der anzeigt, wie 'fuzzy' die Lösung ist. Er liegt immer im Bereich von 1/Anzahl Cluster (vollständig 'fuzzy') bis 1 (vollständig 'crisp, das heißt deutlich getrennt). Zu Grundlagen und genauer Berechnung der genannten Verfahren siehe MATHSOFT, 1997.

In der hierarchische Clusteranalyse dienen neben den Dendrogrammen auch Bannerplots zur Einschätzung der möglichen Anzahl an vorhandenen Gruppen. Heben sich Cluster sehr deutlich voneinander ab, so erscheint dies im Dendrogramm durch sehr kurze Linien bis zum Verschmelzungspunkt von Objekten eines Clusters und sehr lange Linien bis zum Verschmelzungspunkt eines anderen Clusters. Im Bannerplot werden die Verschmelzungspunkte durch horizontale Balken wiedergegeben. Sie beinhalten somit dieselbe Information wie Dendrogramme. Je stärker der Bannerplot durch diese Balken gefüllt bist, desto größer ist die Ähnlichkeit der verschiedenen Cluster, das heißt, desto geringer ist die Clusterstruktur der Gesamtheit der Objekte. Eine zusätzliche Information liefert der agglomerative Koeffizient. Wenn d(i) die mittlere Unähnlichkeit des Objekts i zu dem Cluster ist mit dem es zuerst verschmolzen wird, geteilt durch die Unähnlichkeit dieses Objekts bei der Verschmelzung im letzten Schritt des Clusteralgorithmus, so ist der agglomerative Koeffizient AC definiert als das Mittel aller 1 - d(i). Ein niedriger AC deutet an, daß eine Vergrößerung der Cluster nur zu einer geringen Zunahme der Unähnlichkeiten in diesen Clustern führt, was wiederum ein Indiz für eine recht undeutliche Clusterstruktur ist. Eine analoge Definition gilt für den divisiven Koeffizienten (MATHSOFT, 1997).

BOCK, 1985, nennt alternative Verfahren zur Bestimmung des Vorliegens einer Clusterstruktur. KRZANOWSKI & LAI, 1988, und MILLIGAN & COOPER, 1985, diskutieren die zweite der oben gestellten Fragen, nämlich die Frage nach der optimalen Anzahl an Clustern (wenn denn eine Clusterstruktur überhaupt vorliegt).

b-Multiple Spanning Trees

Multiple Spanning Trees stellen ebenfalls eine Möglichkeit der Repräsentation einer Proximitätsmatrix dar (GOWER & ROSS, 1969). Der Aufbau erfolgt auf iterativem Weg in der Art, daß jedes Objekt durch einen Punkt dargestellt wird, alle Objektpunkte mit Linien verbunden werden, ohne daß geschlossene Verbindungen entstehen, und die Summe der Längen der Verbindungslinien das Minimum aller möglichen Verbindungen darstellt. Die Länge der einzelnen

Liniensegmente entspricht den Werten der Proximitätsmatrix der Objekte. Die Winkel der Verbindungslinien sind in der Regel so zu wählen, daß eine übersichtliche Abbildung entsteht. Allerdings ist auch die Überlagerung des Multiple Spanning Trees über eine zweidimensionale Konfiguration zum Beispiel aus einer Hauptkoordinatenanlyse denkbar (siehe 2.1.2 und Auswertungen in Kapitel 3).

Der Multiple Spanning Tree verdeutlicht, ähnlich wie das Dendrogramm, Objektgruppierungen und visualisiert die Elemente einer Proximitätsmatrix. Er liefert dieselben Objektgruppierungen wie das Dendrogramm der Single-Link-Methode. Für die anderen Clusteranalyseverfahren stellt der Multiple Spanning Tree eine Kontrollmöglichkeit der Angemessenheit bestimmter Gruppenbildungen dar. Schließlich bietet sich die Überlagerung des Multiple Spanning Tree über die Objektabbildung, zum Beispiel einer Hauptkoordinatenanlyse, an. Durch die Dimensionserniedrigung schlecht abgebildete Objekte beziehungsweise Objektdistanzen können durch den Multiple Spanning Tree aufgedeckt werden. Liegen zum Beispiel in einer Hauptkoordinatenanlyse-Abbildung zwei Objekte dicht beieinander, während die Verbindung dieser Objekte im Multiple Spanning Tree nicht auf direktem Weg, sondern über Umwege, das heißt über ein oder mehrere andere Objekte erfolgt, so läßt dies den Schluß auf einer Mißrepräsentation der Objektdistanz in der zweidimensionalen Abbildung zu.

Weiterentwicklungen im Bereich der Multiple Spannung Trees, vor allem auch der Einsatz im Bereich der interaktiven Graphik beschreibt SCHILLER, 1996.

2.5.1.3 Scatterplots und Trellis-Displays

a-Scatterplots

Um die Beziehung zweier Variablen zueinander darzustellen, ist der Scatterplot ein vielfach eingesetztes graphisches Mittel. Er gibt Hinweise auf Beziehungen zwischen den Variablen, auf Gruppierungen bei den Objekten, auf die Verteilung der Werte und auf Ausreißer. Speziell um Beziehungen zwischen Variablen zu verdeutlichen, erfolgt häufig eine Kurvenanpassung an den Punkteschwarm im Scatterplot. Darüber hinaus tragen zu einer effektiven Gestaltung eines Scatterplots das Banking, Jittering und Slicing bei (nach CLEVELAND, 1993).

1. Banking; Banking dient der effektiven Darstellung eines Punkteschwarms beziehungsweise einer an den Punkteschwarm angepaßten Kurve. Eine Kurve mit einer Steigung von 1 besitzt eine Orientierung von 45 Grad, eine Kurve mit einer Steigung von -1 eine Orientierung von - 45 Grad. Eine Zentrierung der absoluten Orientierungen einer Kurve an dieser 45 Grad Linie führt im allgemeinen zur bestmöglichen Wahrnehmung der Kurveneigenschaften; es erfolgt in diesem Fall das sogenannte Banking auf 45 Grad. Banking ergibt je nach Orientierung ein bestimmtes Verhältnis der Y- zur X-Achse, in dem dann die Darstellung des Punkteschwarms mit oder ohne angepaßte Kurve erfolgt. Zur Berechnung siehe CLEVELAND, 1993 und OLLERTON & HARDING, 1995.

2. Jittering; ein Problem, das bei Scatterplots auftreten kann, vor allem bei größeren Datenmengen und ganzzahligen oder stark gerundeten Werte, ist die Überlagerung gleicher Punkte. Unter Jittering versteht man das Hinzufügen einer festgelegten Streuung zu den Werten einer oder beider abgebildeten Variablen. Diese Streuung muß im Vergleich zur Spannweite der Variablenwerte gering sein und in einem auf Null zentrierten Intervall erhoben werden, Häufig bietet sich die Generation von Zufallszahlen ϑ aus der Gleichverteilung an[34].
3. Slicing; unter Slicing versteht man das Betrachten der Werte der einen Variablen an nur einem Wert oder in einem gewählten Intervall der anderen Variablen. Die Werte der ersten Variablen werden dann zum Beispiel in Form eines Boxplots dargestellt oder können für verschiedene Werte beziehungsweise Intervalle mit geeigneten graphischen Mitteln einander gegenübergestellt werden. Entsprechende Intervalle ergeben sich aus dem sogenannten equal-count-Algorithmus, dar nach Vorgabe der Anzahl der Intervalle und der Überlagerung der Intervalle, die Grenzen der Intervalle in der Art liefert, daß jedes Intervall in etwa die gleiche Anzahl an Punkten beinhaltet. Diese Intervalle werden auch als Shingles bezeichnet (zur praktischen Durchführung siehe ebenfalls CLEVELAND, 1993).

Liegt eine weitere, eine dritte Variable vor, so ist eine Scatterplot-Darstellung in drei Dimensionen möglich. Allerdings sind dreidimensionale Scatterplots bei weitem schwerer zu lesen als zweidimensionale. So ist zum Beispiel die Zuordnung von Werten zu einzelnen Punkten im dreidimensionalen Scatterplot recht schwierig. Zu den in Graphikprogrammen üblichen Hilfsmitteln, um auch dreidimensionale Scatterplots besser lesbar zu machen, zählen die Möglichkeiten der Rotation, Farbkodierung, Verknüpfung mit Ausgangsdaten, Rahmengebung und ähnliches.

Häufig übersichtlicher als dreidimensionale Scatterplots sind Scatterplotmatrizen, eine Zusammenfassung aller Scatterplots der drei (oder mehr) betrachteten Variablen in einer Abbildung. Während die Diagonale der Scatterplotmatrix die Variablenbenennungen enthält, sind die einzelnen Scatterplots aller Variablenpaare sowohl oberhalb als auch unterhalb der Diagonalen abgebildet. Die Scatterplots sind mit entsprechenden Skalen und Referenzlinien zu versehen, um die Les- und Interpretierbarkeit zu verbessern. Die Inspektion einer Scatterplotmatrix kann darüber hinaus durch interaktive, graphische Instrumente vertieft werden.

b-Trellis-Displays

Trellis-Displays, die auch als Co-Plots (conditioning plots) bezeichnet werden, erweitern noch die Möglichkeiten der Scatterplotmatrizen, mehrdimensionale Sachverhalte in einer Abbildung

---

[34] $\vartheta \sim U(a, b)$ mit Grenzen a und b $(b > a)$ und $E(\vartheta) = (a + b) / 2 = 0)$.

aufzuzeigen. Die Grundlagen werden von BECKER et al., 1994, und CLEVELAND, 1993, erläutert. THEUS, 1996, vergleicht Trellis-Displays und interaktive Graphik.

Es handelt sich bei Trellis-Displays um eine nach einem bestimmten Schema aufgebaute Anordnung von Einzelgraphiken, die nach THEUS, 1996, Informationen von bis zu acht Variablen auf einer (DIN A4) Seite darstellen können. Zur Spezifikation eines Trellis-Displays gehört die Angabe der verwendeten Daten, die graphische Methode der Einzelgraphiken (zum Beispiel Scatterplots, Linienplots, Boxplots), die Benennung der zwei- (oder drei) Achsenvariablen und die Benennung der konditionierenden Variablen. Die konditionierenden Variablen können nominal- oder ordinalsklierte Variablen, oder intervall- beziehungsweise verhältnisskalierte Variablen sein, die zuvor in Klassen eingeteilt werden, zum Beispiel nach Maßgabe des equal-count-Algorithmus. Durch die konditionierenden Variablen können jeweils Bereiche für diese Variablen festgelegt werden, in denen die Werte der Achsenvariablen abgebildet werden sollen. Die Definition der Variablen als konditionierende Variablen und Achsenvariablen ist im Prinzip beliebig und kann daher zu einer Vielzahl von Trellis-Displays mit jeweils unterschiedlichen Festlegungen führen. Je größer die Variablenzahl, desto schwieriger wird es einen Gesamteinblick, in allen möglichen Variablenkombinationen zu bekommen. Die Klassenbildung bei intervall- oder verhältnisskalierten Variablen unterliegt hier, wie auch in anderen Methoden, zum Beispiel der Korrespondenzanalyse, einer gewissen Subjektivität.

Referenzlinien, die nicht mit bestimmten Achsenwerten übereinstimmen müssen, können hilfreich sein, um Werte zwischen Einzelgraphiken zu vergleichen.

Ein Verfahren, das im Auswertungteil häufig eingesetzt wird, ist die Anpassung sogenannter Loess-Linien an einen Punkteschwarm in den Panels eines Trellis-Displays. Die Loess-Linien werden an Stelle parametrischer Regressionslinien (zum Beispiel einer linearen Kleinste-Quadrate-Regression) gewählt, da sie weniger stark auf Extremwerte und Ausreißer, reagieren. Loess steht für local regression und wird in CLEVELAND, 1993, beschrieben. An Stelle einer einmaligen Kurvenanpassung an alle Werte, erfolgt eine schrittweise, lokale Kurvenanpassung im Bereich jedes einzelnen Punktes, unter Berücksichtigung des Gewichts der ihn umgebenden Punkte, wobei näherliegende Punkte ein höheres Gewicht haben als weiter entfernt liegende. Die lokale Anpassung erfolgt so mit einer gewichteten linearen oder quadratischen Kleinste-Quadrate-Regression und ergibt einen Loess-Schätzer für den gewählten Punkt. Dieser Vorgang wird für alle Punkte wiederholt. Die Loess-Schätzer werden dann durch Liniensegmente miteinander verbunden. Unterschiede in der Kurvenanpassung ergeben sich durch die Festlegung der Loess-Parameter, das heißt des Glättungsparameters, der festlegt wie groß der Bereich von Punkten ist, der in der lokalen Anpassung betrachtet werden soll[35], und des Regressionsparameters, der bestimmt, ob

[35] Der Glättungsparameter liegt in der Regel zwische 0,25 und 1. Ein Glättungsparameter von 0,5 bedeutet zum Beispiel, daß bei Vorliegen von 20 Werten, 10 Werte zur lokalen Anpassung

eine lokal lineare oder eine lokal quadratische Anpassung erfolgen soll. Schließlich können die Residuen zwischen beobachteten und geschätzten Werten noch in die Berechnung miteinbezogen werden wodurch dann auf iterativen Weg eine Minimierung der Residuen erreicht wird. Dieses Verfahren führt zu einer erhöhten Robustheit bei Vorliegen extremer Werte.

2.5.1.4 Interaktive Graphik und sonstige Verfahren

Der Einsatz interaktiver Graphikprogramme wie zum Beispiel Manet oder Data Desk, ist eine wichtiger Schritt im Bereich der Datenanalyse, um Daten kennenzulernen und zu hinterfragen, beziehungsweise um entsprechende Hypothesen zu den Daten zu entwickeln. UNWIN, 1992, und THEUS, 1996, vermuten eine ständig zunehmende Bedeutung interaktiver Graphiken für die statistische Datenanlyse. Wesentliche Bestandteile interaktiver Graphikprogamme sind (CLEVELAND, 1993, NAGEL et al., 1992, OSTERMANN & WOLF-OSTERMANN., 1992):

- Rotationsmöglichkeiten;
- Identifikation von Objektbenennungen, Variablenwerten, Achsenskalen und ähnlichem;
- Isolation einzelner Variablen und Objekte;
- Verknüpfung gleicher Objekte beziehungsweise Variablen in unterschiedlichen Graphiken;
- Markierung oder Maskierung von Objekten oder Variablen;
- selektive Benennung einzelner Objekte (brushing);
- unterschiedliche graphische Darstellungsmöglichkeiten (Scatterplots, Histogramme, Boxplots, und andere mehr).

Die zunehmende Leistungsgfähigkeit von Computern wird auch die Leistungsfähigkeit interaktiver graphischer Verfahren noch steigern. Die Faszination interaktiver Graphiken läßt sich jedoch nur schwer auf statische Dokumente wie Bücher übertragen. Daher sind sie zur Vermittlung von Untersuchungsergebnissen auf die Darstellung am Computer beschränkt.

Zur Darstellung mehrdimensionaler Sachverhalte im üblichen, zweidimensionalen Publikationsformat geben TUKEY & TUKEY, 1981, eine Vielzahl weiterer Hinweise. Dazu zählen die geeignete Auswahl von Symbolen, der Verzicht auf die Darstellung bestimmter Punkte und die Verwendung von Kontur- und Referenzlinien. Darüber hinaus sind Agglomeration von Punkten ähnlicher Werte möglich. Denkbar ist auch die Aufteilung einer Abbildung in viele einzelne Bereiche, die dann mit zusätzlichen Informationen zu den Werten in diesem Bereich gefüllt werden können (sogenannte multiwindow plots). Auch können die Symbole zweier Variablen mit weiteren Informationen zum Beispiel über eine dritte Variable versehen werden (durch Linien, Polygone,

---

ausgewählt werden, und zwar die 10, die dem Wert, für den der Schätzer berechnet werden soll, am nächsten liegen.

Farben, Zeichenstärke und ähnliches). Eine Anzahl von Variablen lassen sich auch als eigene Symbole darstellen, wobei die Gestaltung des Symbols von den Werten der Variablen abhängt. Hierzu zählen zum Beispiel die Starplots, Cherneyeff Gesichter, Andersons Glyphen, Kleiner-Hartigan-Bäume und viele andere mehr. Farbplots (Dshade-Plots) schließlich können Variablenwerte unterschiedlicher Objekte oder die Werte von Proximitätsmatrizen durch unterschiedliche Farben visualisieren.

Darstellungsmöglichkeiten gibt es demnach in großer Zahl. Inwieweit einzelne Vorgehensweisen tatsächlich das Verständnis für die Daten vertiefen beziehungsweise mehrdimensionale Sachverhalte in einfachen Abbildungen zusammenfassen, ist nicht grundsätzlich zu beantworten. Vielmehr sind für den Einzelfall je nach Fragestellung sowie Objekt- und Variablenzahl geeignete Darstellungsformen zu wählen.

2.5.2 Ergänzende Methoden

In den folgenden Kapiteln werden vier Bereiche kurz angesprochen, die von allgemeiner Bedeutung in der multivariaten Datenanalyse sind.

Der Überprüfung von Daten auf Vorliegen der Multinormalverteilung und der Varianzhomogenität kommt vor allem Bedeutung beim Einsatz schließender Verfahren zu (2.5.2.1). Die Verwendung robuster Methoden spielt in erster Linie dort eine Rolle, wo einzelne, untypische Objekte (Ausreißer) einen starken Einfluß auf die Lösung haben (2.5.2.2). Eng verknüpft mit der Thematik der Ausreißer ist die Frage nach dem Umgang mit fehlenden Werten in multivariaten Datensätzen (ebenfalls 2.5.2.2). Schließlich ist die Stabilität einer Analyselösung mit geeigneten Verfahren zu überprüfen (2.5.2.3).

2.5.2.1 Tests auf Multinormalverteilung und Varianzhomogenität

a-Multinormalverteilungstests

Normalverteilungstests lassen sich im univariaten Fall nach KOZIOL, 1986, in vier Gruppen einteilen[36]:

1. Tests der Anpassungsgüte nach Shapiro-Wilk und Abwandlungen davon;
2. Tests, die auf dem Vergleich mit der empirischen Verteilungsfunktion aufbauen (zum Beispiel Kolmogorov-Smirnov);
3. Berechnung und Beurteilung von Schiefe und Kurtosis;
4. informelle, graphische Methoden.

Für den multivariaten Fall gibt es nun verschiedene Generalisierungen für die univariaten Verfahren.

So schlägt ROYSTON, 1983, die Inspektion der Shapiro-Wilk-Statistik für jede einzelne Variable vor und beschreibt ein Verfahren der Kombination der einzelnen Werte, um eine Aussage zur Multinormalverteilung zu treffen.

Auch auf der empirischen Verteilungsfunktion beruhende Verfahren lassen sich für den multivariaten Fall konstruieren, ihre praktische Bedeutung ist aber gering.

Mit der Berechnung von Schiefe und Kurtosis und der Erarbeitung aussagekräftiger Statistiken für die multivariable Fragestellung beschäftigen sich zum Beispiel MACHADO, 1983, MALKOVICH & AFIFI, 1973, und SMALL, 1980.

---

[36] Es ist anzumerken, daß für den Fall, daß Multinormalverteilung zutrifft, gilt, daß alle Variablen univariat normalverteilt sein müssen, daß aber die univariate Normalverteilung aller Variablen, allein noch kein ausreichender Hinweis auf das Vorliegen der Multinormalverteilung ist.

Daneben gibt es Multinormalverteilungstests, die kein univariates Gegenstück besitzen. Hierzu zählen die von GNANADESIKAN, 1977, ausführlich beschriebenen Winkel- und Radientests. Die Einzelheiten der verschiedenen Verfahren sind den angegebenen Quellen zu entnehmen.

Ist Multinormalverteilung nicht gegeben und sollen auf der Annahme der Multinormalverteilung beruhende Verfahren eingesetzt werden, bietet sich eine entsprechende Transformation der Variablen an. ANDREWS et al., 1971, beschreiben Möglichkeiten zur Transformation zur Multinormalverteilung. Eine Transformation einzelner, nicht normalverteilter Variablen in normalverteilte transformierte Variablen, reicht in der Regel nicht aus, um zur Multinormalverteilung zu gelangen. Das Vorgehen kann aber zumindest zu einer symmetrischen Verteilung führen. Nach KRZANOWSKI, 1988a, ist die Anwendung schließender, auf der Annahme der Multinormalverteilung beruhender, Verfahren in vielen Fällen möglich, solange die Werte zumindest aus einer zentral symmetrischen Verteilung stammen.

b-Varianzhomogenitätstests

Liegen Daten gruppiert vor, so ist die Frage zu stellen, ob allen Gruppen eine gemeinsame Kovarianzmatrix zugrunde liegt. Die Überprüfung der Varianzhomogenität spielt zum Beispiel in der Diskriminanzanalyse oder in der multivariaten Varianzanalyse eine große Rolle, weniger allerdings in den Verfahren, die überwiegend in dieser Arbeit eingesetzt werden. Die multivariate Variante des Bartlett Tests, der allerdings auch sehr stark auf Abweichnungen von der Multinormalverteilung reagiert, stellt eine Möglichkeit dar, die Gleichheit mehrerer Kovarianzmatrizen zu testen (siehe zum Beispiel HAND & CROWDER, 1996). Ein graphisches Verfahren, aufbauend auf der Biplot-Methodologie, stellen CORSTEN & GABRIEL, 1976, vor.

Zusätzlich kann es zum Beispiel in der Hauptkomponentenanalyse wichtig sein zu klären, ob die einzelnen Variablen in etwa die gleiche Variabilität aufweisen, um eine Entscheidung bezüglich der Notwendigkeit einer Standardisierung der Variablen zu treffen. Verschiedene bekannte Verfahren, die die Homogenität von Varianzen prüfen sind neben dem Bartlett Test der Box-Scheffe-, Levene-, F-, oder Cochran-Test (RASCH et al., 1992, SOKAL & ROHLF, 1981). Diese Tests gehen allerdings vom Vergleich von Varianzen von voneinander unabhängigen Behandlungen aus. Die Überprüfung der Varianzhomogenität in Datensätzen mit untereinander korrelierten Variablen, wie sie in der dieser Arbeit vorliegen, wird zum Beispiel von HARRIS, 1985, besprochen. Er entwickelt vier Teststatistiken W, W_R, W_L und W_G. Im Fall von W_L und W_G wird mit der Logarithmus-Transformation gearbeitet und somit eine gewisse Stabilität gegenüber Abweichungen von der Multinormalverteilung, so sie denn vorliegt, Rechnung getragen. In großen Stichproben folgen die Teststatistiken der Chi-Quadrat-Verteilung mit $p-1$ Freiheitsgraden.

2.5.2.2 Robuste Methoden und fehlende Werte

a-Robuste Methoden

Häufig tauchen in einem Datensatz zweifelhafte oder untypische Werte auf. Diese können durch falsche Messungen verursacht sein, oder auf Übertragungsfehlern, einer falschen Kommasetzung oder ähnlichem beruhen. Natürlich kann es sich auch um tatsächlich extreme Werte handeln, die zwar richtig aufgezeichnet sind, aber eben aus dem allgemeinen Rahmen der übrigen Werte fallen. Wie richtig mit derartigen Ausreißern umzugehen ist, ist nicht unumstritten. SEBER, 1984, führt verschiedene Standpunkte aus. Eine Möglichkeit - die einzige, die an dieser Stelle vertieft wird - ist die Erarbeitung sogenannter robuster Schätzer beziehungsweise der Einsatz robuster Verfahren. Ein guter robuster Schätzer, zum Beispiel für die Kovarianzmatrix, weist bei einem der Multinormalverteilung folgenden Datensatz ohne untypische Werte eine hohe Effizienz gegenüber des besten unverzerrten Schätzers (also dem Maximum Likelihood Schätzer **S** für $\mathbf{\Sigma}$) auf. Liegen untypische, zweifelhafte Werte vor, so wird der robuste Schätzer weniger stark durch diese beeinflußt als der nicht robuste Schätzer.

Im Fall der Hauptkomponentenanalyse kann es daher zum Beispiel angebracht sein, an Stelle der Kovarianzmatrix **S** eine robuste Variante der Kovarianzmatrix zu verwenden. Verschiedene Vorschläge, wie man zu einem robusten Schätzer der Kovarianzmatrix gelangt, geben JACKSON, 1991, KRZANOWSKI & MARRIOTT, 1994, oder SEBER, 1984. Die Methode von CAMPBELL, 1980, soll kurz ausgeführt werden. Untypische Werte werden derart gehandhabt, daß ihnen ein geringes Gewicht w_i zufällt. Erhalten alle Objekte das Gewicht $w_i = 1$ so werden alle Objekte gleich und voll gewichtet. Erhält eine Objekt zum Beispiel das Gewicht $w_i = 0$, so wird dieses Objekt überhaupt nicht berücksichtigt. Je untypischer ein Objekt ist, desto geringer ist w_i im Bereich von 0 bis 1. Es ergeben sich dann mit den Gewichten als robuste Schätzer für den Mittelwertsvektor $\bar{\mathbf{x}}_\mathbf{M}$ und die Kovarianzmatrix $\mathbf{S}_\mathbf{M}$, $\bar{\mathbf{x}}_\mathbf{M} = \sum_{i=1}^{n} w_i \mathbf{x_i} \,/\, \sum_{s=1}^{n} w_i$ und

$\mathbf{S}_\mathbf{M} = \sum_{i=1}^{n} w_i^2 (\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{M})(\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{M})' / (\sum_{i=1}^{n} w_i^2 - 1)$. Die Gewichte errechnen sich auf iterativem Weg. Es erfolgt zunächst die Berechnung der Mahalanobis-Distanz dm_i für Objekt i vom Zentroid nach $dm_i = (\mathbf{x_i} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}})$ und daraus dann $w_i = \psi(dm_i) / dm_i$, mit
$\psi(dm_i) = dm_i$, wenn $dm_i \le dm_0$ beziehungsweise
$\psi(dm_i) = dm_0 e^{-1/2} (dm_i - dm_0)^2 / \beta$, wenn $dm_i > dm_0$, wobei für dm_0 gilt:
$dm_0 = \sqrt{p} + \alpha\sqrt{2}$. Mit den Gewichten werden dann Zentroid und Kovarianzmatrix, sowie die daraus resultierende Mahalanobis-Distanz erneut berechnet und dieser Vorgang bis zur Konvergenz wiederholt. Für den Iterationsprozeß sind nun noch die Parameter α und β festzulegen. CAMPBELL, 1980, empfiehlt die folgenden Varianten:

1. $\alpha = \infty$ (β dann unbedeutend); dies führt zu $w_i = 1$ für alle i.
2. $\alpha = 2$; $\beta = \infty$; dies führt zu $w_i = 1$, wenn $dm_i \le \sqrt{p} + \sqrt{2}$ und zu
   $w_i = (\sqrt{p} + \sqrt{2}) / dm_i$, wenn $dm_i > \sqrt{p} + \sqrt{2}$.

3. $\alpha = 2$, $\beta = 1{,}25$; dies führt zu einer Vielzahl von Gewichten, entsprechend der Werte der einzelnen Objekte.

Die Identifikation der untypischen Objekte erfolgt zum Beispiel durch die entsprechenden, geringen Gewichte. Als Faustzahlen für untypische Werte nennt CAMPBELL, 1980, ein Gewicht von < 0,3 bei $\alpha = 2$ und $\beta = 1{,}25$. Demgegenüber ist ein Gewicht von > 0,7 bei $\alpha = 2$ und $\beta = 1{,}25$ ein Indiz für eine typische, nicht aus dem Rahmen fallende Einheit. Eine weitere Möglichkeit ist die Erstellung einer Graphik mit den der Größe nach geordneten Werten für $dm_i^{2/3}$ gegen die Quantile der Normalverteilung. Untypische Objekte können wie im normalen q-q-Plot auch durch Abweichungen vom linearen Verlauf erkannt werden.

Für die ordinale mehrdimensionalen Skalierung schildern SPENCE & LEWANDOWSKI, 1989, ein robustes Vorgehen, wobei es hier jedoch nicht um die Berechnung eines robusten Schätzers für die Proximitätsmatrix, sondern um die Anwendung eines robusten Algorithmus geht. Mit dem Programm TUFSCAL kann diese robuste Version der ordinalen mehrdimensionalen Skalierung durchgeführt werden.

b-Fehlende Werte

Treten fehlende Werte auf, so ist eine Möglichkeit ist der Verzicht auf alle Objekte, die bei einer Variablen einen fehlenden Wert aufweisen. Dies kann jedoch unter Umständen dazu führen, daß ein Großteil der vorhanden Informationen verworfen wird.

Eine zweite, ebenfalls unbefriedigende Lösung, ist die Verwendung der jeweiligen Mittelwerte der einzelnen Variablen an Stelle der fehlenden Werte. Insbesondere bei der Berechnung von Proximitätsmatrizen warnt KRZANOWSKI, 1988a, vor einem einfachen Ersatz der fehlenden Werte durch die Variablenmittelwerte, da diese fast immer zu einer Unterschätzung der tatsächlichen Proximitäten führen. Besser ist es dann schon für die Berechnung der Proximität zwischen zwei Objekten nur die Variablen ohne fehlende Werte zu verwenden, das heißt nur q von p Variablen zu verwenden ($q < p$), und dann den errechneten Wert mit dem Faktor p / q zu multiplizieren.

Daneben gibt es verschiedene Vorgehensweisen, die auf iterativem Weg einen geeigneten Wert für den fehlenden Wert suchen. Unter der Annahme der Multinormalverteilung ist zum Beispiel die Methode nach BEALE & LITTLE, 1975, zu nennen. Ohne Verteilungsannahmen kommt die direkt mit der Eigenwertzerlegung der Ausgangsmatrix arbeitende Methode von KRZANOWSKI, 1988b, aus. Einzelheiten dieser, oder anderer Möglichkeiten werden hier nicht angesprochen.

2.5.2.3 Beurteilung der Stabilität

Eine Analyse wird als stabil bezeichnet, wenn ‘geringe’ Veränderungen in den Daten zu ‘geringen’ Veränderungen bei den Ergebnissen führen. GIFI, 1990, unterscheidet verschiedene Formen der Stabilität, so die analytische und algebraische sowie die Wiederholungsstabilität. Daneben weist er auf Stabilitätsgesichspunkte bezüglich der Daten- und Modellselektion hin. GREENACRE, 1984,

stellt interne Stabilität (Ausreißer) der externen Stabilität (Wiederholungsstabilität) gegenüber (siehe 2.1.3).

Zwei Verfahren, die zur Überprüfung der Stabilität eingesetzt werden können, sind das Jackknifing und das Bootstrapping (SHAO & TU, 1996). Im Fall des Jackknifinig werden n Analysen mit jeweils $n-1$ Objekten durchgeführt, das heißt nacheinander werden die Ausgangsdaten ohne die erste, dann ohne die zweite, dann ohne die dritte Einheit und so weiter analysiert. Beim Bootstrapping hingegen wird aus den Ausgangsdaten k-mal eine Stichprobe vom Umfang n mit Zurücklegen gezogen. Da zurückgelegt wird und alle Objekte die gleiche Wahrscheinlichkeit $1/n$ haben in die Bootstrap-Stichprobe zu gelangen, werden in der Regel einige Objekte häufiger vertreten sein als andere.

Bei Verwendung von Hauptkomponenten- oder Korrespondenzanalyse können die Jackknife oder Bootstrap Lösungen, das heißt die bei den einzelnen Wiederholungen errechneten Ergebnisse der Objektwerte im dimensionserniedrigten Raum (in zwei Dimensionen), in die Konfiguration der Analyse aller Werte der Ausgangsmatrix projiziert werden. Durch Linien, die die äußersten Objektwerte eines Objekts miteinander verbinden, ergeben sich dann je Objekt konvexe Hüllen, die bei Bedarf noch ‘geschält’ werden müssen, um auf instabile Jackknife- oder Bootstrap-Stichproben zu reagieren (GREEN, 1981). Im Fall von Optimierungsverfahren ist eine einfache Projektion neuer Objektwerte in die ursprüngliche Objektkonfiguration nicht möglich. Allerdings können die Ergebnisse der wiederholten Analysen mit Hilfe der Prokrustes-Analyse miteinander verglichen werden. Dies empfiehlt sich auch dann, wenn viele Objekte vorliegen, da die Überlagerung einer Vielzahl von Stichproben letztlich zu einer großen Unübersichtlichkeit führen würde.

Daneben können Jackknife- und Bootstrap-Schätzer der Korrelations- oder Kovarianzmatrix auch den Ausgangspunkt einer Hauptkomponentenanalyse oder Faktoranalyse darstellen und somit eine robuste Alternative zu dem unter 2.5.2.2 vorgestellten Verfahren bieten.

Tabelle 6: Agglomerationskriterien unterschiedlicher Clusterverfahren

Clusterverfahren	Agglomerationskriterum
Single-Link-Methode	Zwei Objekte oder Objektgruppierungen werden zu einer neuen Gruppierung zusammengefaßt, wenn die kleinste Unähnlichkeit zwischen zwei Objekten unterschiedlicher Objektgruppierungen (wobei anfänglich eine Objektgruppierung auch aus einem einzigen Objekt bestehen kann), das Minimum aller Unähnlichkeiten zwischen allen Objektgruppierungen darstellt.
Complete-Link-Methode	Zwei Objekte oder Objektgruppierungen werden zu einer neuen Gruppierung zusammengefaßt, wenn die größte Unähnlichkeit zwischen zwei Objekten unterschiedlicher Objektgruppierungen (wobei anfänglich eine Objektgruppierung auch aus einem einzigen Objekt bestehen kann), das Minimum aller Unähnlichkeiten zwischen allen Objektgrupppierungen darstellt.
Group-Average-Methode	Zwei Objekte oder Objektgruppierungen werden zu einer neuen Gruppierung zusammengefaßt, wenn die mittlere Unähnlichkeit zwischen zwei Objekten unterschiedlicher Objektgruppierungen (wobei anfänglich eine Objektgruppierung auch aus einem einzigen Objekt bestehen kann), das Minimum aller Unähnlichkeiten zwischen allen Objektgrupppierungen darstellt.
Zentroid-Methode	Zwei Objekte oder Objektgruppierungen werden zu einer neuen Gruppierung zusammengefaßt, wenn die quadrierte euklidische Distanz zwischen zwei Objekten unterschiedlicher Objektgruppierungen (wobei anfänglich eine Objektgruppierung auch aus einem einzigen Objekt bestehen kann), das Minimum aller Unähnlichkeiten zwischen allen Objektgrupppierungen darstellt.
Median-Methode	wie die Zentroid-Methode. Allerdings wird bei der Neuberechnung der Variablenwerte einer entstandenen Objektgruppierung nicht der Mittelwert (wie in allen übrigen Methoden auch), sondern der Median verwendet.
Minimum-Variance-Methode	Zwei Objekte oder Objektgruppierungen werden zu einer neuen Gruppierung zusammengefaßt, wenn die Zunahme der Varianz einer Objektgruppierung durch Hinzunahme einer weiteren Objektgruppierung das Minimum aller möglichen Agglomerationen darstellt.

3 Beispiele der Visualisierung und Analyse

3.1 Betriebsbegleitende Untersuchung der Cyclamenkultur in 20 westfälisch-lippischen Gartenbaubetrieben von 1994

3.1.1 Einführung

Betriebsbegleitende Untersuchungen spielen für die Beratung im Gartenbau eine große Rolle. Allein an der Bezirksstelle für Gartenbau der Landwirtschaftskammer Westfalen-Lippe in Münster wurden seit 1993 derartige Untersuchungen bei Poinsettien, Primeln, Hortensien (Rohwarekultur und Treiberei) und Cyclamen durchgeführt (JOKIEL, 1996). Aus dem Obstbau liegt ein Beispiel von GÖRGENS, 1991, vor. BOKELMANN, 1987, entwickelt, aufbauend auf den Ergebnissen und Beobachtungen einer betriebsbegleitenden Untersuchung bei Poinsettien, Grundlagen von Entscheidungsprozessen im Gartenbaubetrieb. WIESMANN, 1985, beleuchtet die Relevanz derartiger Untersuchungen für die Arbeit der Beratung.

Neben der Möglichkeit, Erkenntnisse zum Produktionsablauf zu gewinnen, bietet eine betriebsbegleitende Untersuchung der Beratung die Möglichkeit engen Kontakt zu den beteiligten Betrieben aufzubauen und auch den fachlichen Austausch der Betriebsleiter untereinander zu fördern. Insofern erfüllt eine betriebsbegleitende Untersuchung auch eine nicht zu unterschätzende ‘soziale’ Komponente.

Die Darstellung des in der Regel sehr umfangreichen Datenmaterials ist allerdings in vielen Beispielen unbefriedigend, und zwar in dem Sinne, daß es kaum gelingt aus den tabellarischen und graphischen Darstellungen einen zusammenfassenden Überblick über die wirklich wesentlichen und auffälligen Informationen zu gewinnen (siehe zum Beispiel BOKELMANN & VOTH, 1986, BLECKEN, 1987, JOKIEL & HOCKWIEN, 1994).

Mit Hilfe der im Kapitel 2 besprochenen Methoden soll nun der Einblick in die Daten einer typischen betriebsbegleitenden Untersuchung, wie sie an der Landwirtschaftskammer Westfalen-Lippe durchgeführt werden, vertieft werden. Dabei interessieren vor allem die folgenden Fragen:

1. Welche Ähnlichkeitsbeziehungen bestehen zwischen den Betrieben hinsichtlich der ausgewählten und bestimmten Merkmale?
2. Weichen die Ähnlichkeitsbeziehungen zwischen den Betrieben in einzelnen Variablensets stark voneinander ab?
3. Läßt das Kulturverfahren einen Rückschluß auf die Produktqualität zu?
4. Welche Korrelationen bestehen zwischen den Variablen gleicher und unterschiedlicher Variablensets?

Grundlegende Probleme der Daten stellen zum einen die Frage nach der Aufzeichnungsgenauigkeit und zum anderen die insgesamt geringe Anzahl an Betrieben, die sich an der Untersuchung beteiligt haben, dar. Insofern läßt die betriebsbegleitende Untersuchung keine

Rückschlüsse auf die Cyclamenkultur im Allgemeinen zu. Die im den folgenden Kapiteln dargestellte Auswertung dient somit in erster Linie der Darstellung der Daten, die zu einer verbesserten Kommunikation der in den Daten enthaltenen Informationen führt. Im Vordergrund steht das Bemühen um eine kompakte und prägnante Darstellung der durch die betriebsbegleitende Untersuchung gewonnenen Informationen.

Die betriebsbegleitende Untersuchung, die nun besprochen wird, wurde 1994 durchgeführt. Beteiligt waren 20 Betriebe, die zum gleichen Zeitpunkt ein einheitliches Cyclamen-Ausgangsmaterial der Sorten ‘Sierra’ und ‘Concerto’ gestellt bekamen und dann im Laufe der Kultur Aufzeichnungen zum Kulturablauf, zur Klimaführung und zu Düngung und Pflanzenschutz erstellten. Zusätzlich wurden Strukturdaten der Betriebe erhoben und an drei Zeitpunkten Substratanalysen durchgeführt. Im Anschluß an die Kultur in den Betrieben erfolgte eine Haltbarkeitsprüfung am Bildungs- und Versuchszentrum des Gartenbaus, Wolbeck, während der zu drei Zeitpunkten Qualitätsbonituren durchgeführt wurden. Der genaue Ablauf der Untersuchung ist JOKIEL & HOCKWIEN, 1994, und PETERS, 1994, zu entnehmen. Die Übersichten A1 bis A4 geben einen Überblick über die in der Untersuchung erfaßten Merkmale.

3.1.2 Darstellung der Ergebnisse

Nacheinander werden nun die vier Variablensets ausgewertet und dann einer Gesamtbetrachtung unterzogen. Es wird so vorgegangen, daß zunächst eine oder mehrere Abbildungen kurz erläutert werden (gekennzeichnet durch vor den Absatz gestelltes „ – “-Zeichen), und dann interpretiert werden (gekennzeichnet durch vor den Absatz gestelltes „ ⇒ “-Zeichen). Spezielle methodische Anmerkungen folgen auf ein „ ∗ “-Zeichen. Die Abbildungen und Übersichten sind im Anhang Teil I A und II A hinterlegt.

3.1.2.1 Variablenset 1 - Beurteilung der Qualität

a-Einführende Datenanalyse

– Einen ersten Einblick in die Qualitätsbonituren der beteiligten Betriebe geben die farbcodierten Starplots in Abbildung A1. Die Länge der Strahlen entspricht dem Boniturwert, rote Strahlen liegen im Bereich von 1 bis 3, grüne Strahlen im Bereich von 4 - 6 und blaue Strahlen im Bereich von 7 - 9. Jeder Strahl entspricht einer der 17 Variablen (fünf Merkmale an drei Zeitpunkten und ein Merkmal (Wurzelqualität) an zwei Zeitpunkten bestimmt), jeder Stern einem Betrieb.

⇒ Besonders auffällig ist das schlechte Ergebnis von Betrieb 3, vor allem bei der Sorte ‘Sierra’, und von Betrieb 19, vor allem bei der Sorte ‘Concerto’. Demgegenüber fallen die Betriebe 2, 11, 12 und 13 bei ‘Sierra’ und die Betriebe 1, 5 und 13 bei ‘Concerto’ durch relativ gute Ergebnisse auf. Neben der Betrachtung einzelner Betriebe ist aber der grundsätzliche Eindruck festzuhalten, daß die Beurteilungen bei ‘Sierra’ insgesamt gesehen etwas besser zu sein scheinen als bei ‘Concerto’, und daß eine Betrieb-Sorte-Wechselwirkung vermutet werden kann, daß also Betriebe mit recht guten Ergebnissen bei einer Sorte durchaus weniger gute Ergebnisse bei der anderen Sorte haben können (siehe zum Beispiel die Betriebe 10, 11 und 12 (‘Sierra’ besser als ‘Concerto’), beziehungsweise die Betriebe 1, 5, und 14 (‘Concerto’ besser als ‘Sierra’)).

– Zur Veranschaulichung, ob ‘Sierra’ tatsächlich in der Regel etwas besser beurteilt wurde als ‘Concerto’, und um zu zeigen, wie sich die Qualitätsbeurteilungen im Mittel mit der Zeit verändert haben, dient der Dotplot in Abbildung A2. Er zeigt die Mediane (über alle Betriebe) aller Merkmale. Der Kreis steht für ‘Sierra’, das Kreuz für ‘Concerto’.

⇒ Man sieht, daß die mittlere Beurteilung sich bei allen Merkmalen mit der Zeit verschlechtert. ‘Sierra’ wird im Mittel bei allen Merkmalen, außer beim Wurzelbild in Woche 44, immer gleich oder besser als ‘Concerto’ beurteilt. ‘Concerto’ rutscht in der Schlußbewertung in Woche 48 bei immerhin drei Merkmalen (Gesamteindruck, Knospenbesatz und Welke) im Mittel unter die Boniturnote 5, den Mittelpunkt der gewählten Ordinalskala.

– Die vier Trellis-Displays der Abbildung A3 beinhalten die einzelnen Boniturwerte der teilnehmenden Betriebe für die Bewertungswochen 44 und 48. Sie sind in der Art geordnet, daß in jedem Trellis-Display im Panel unten links der Betrieb mit dem kleinsten Median (über alle

Merkmale), also der insgesamt am schlechtesten beurteilte Betrieb steht, und daß oben rechts der Betrieb mit dem höchsten Median (über alle Merkmale), also der insgesamt am besten beurteilte Betrieb steht. Die Merkmale sind geordnet nach: Gesamteindruck (1), Knospenbesatz (2), Wurzelbild (3), Fehlen von Vergilbung (4), Fehlen von Welke (5), Fehlen von Krankheiten (6).

⇒ In Woche 44 wird bei beiden Sorten die Ware von Betrieb 13 am besten, die von Betrieb 3 am schlechtesten beurteilt. Die Beurteilungen bei ‘Concerto’ erscheinen insgesamt etwas gleichmäßiger zu sein als bei ‘Sierra’. In Woche 48 fällt vor allem bei ‘Concerto’ die Zunahme der Variabilität der Bonituren zwischen und innerhalb der einzelnen Betriebe und ein etwas schlechteres Ergebnis im Vergleich zu ‘Sierra’ auf (zum Beispiel beim Vergleich der besten Betriebe, also in den obersten Zeilen der jeweiligen Trellis-Displays für Woche 48). Interessant ist auch die Entwicklung von Betrieb 3. Während er bei ‘Sierra’ zu beiden Zeitpunkten an unterster Stelle liegt, ist er bei ‘Concerto’ in Woche 48 bis ins Mittelfeld der Betriebe aufgerückt. Verschiebungen in der Rangfolge der Betriebe sind in allen Abbildungen deutlich, ebenso wie die schon in den Starplots auffällige Betriebe-Sorte Wechselwirkung.

– Eine andere Trellis-Darstellung wird für die beiden Merkmale Gesamteindruck und Knospenbesatz in den Trellis-Displays in den Abbildungen A4 und A5 gewählt (für die übrigen Merkmale ergeben sich ähnliche Abbildungen). Die Sortierung der Panels basiert wieder auf dem Median, so daß links unten Woche und Sorte mit dem niedrigsten Median des Merkmal (über alle Betriebe) und analog rechts oben Woche und Sorte mit dem höchsten Median steht. Die Betriebe sind ihrerseits nach ansteigendem Median über alle Sorte-Woche-Kombinationen sortiert. Die rote Referenzlinie grenzt den Bereich niedriger gegen den Bereich mittlerer Boniturwerte, die blaue Referenzlinie den Bereich mittlerer gegen den Bereich hoher Boniturwerte ab.

⇒ ‘Sierra’ wird demnach bei beiden Merkmalen im Mittel besser beurteilt als ‘Concerto’. Bei beiden Sorten nimmt die Beurteilung von Woche 44 über Woche 46 zu Woche 48 ab. Die stärkere Streuung der Boniturwerte zwischen den Betrieben bei ‘Sierra’ fällt insbesondere beim Gesamteindruck in Woche 48 auf. Beim Knospenbesatz findet bei ‘Concerto’ eine auffällig stärkere Verschlechterung von Woche 44 zu Woche 48 statt als bei ‘Sierra’. Schließlich ist anzumerken, daß die Rangfolge der Betriebe beim Knospenbesatz deutlich von der Rangfolge der Betriebe beim Gesamteindruck abweicht. Die Korrelationsmatrizen mit den Spearman-Rangkorrelations-koeffizienten in den Übersichten A5 und A6 zeigen, daß dies kein Einzelfall ist, sondern (erstaunlicherweise) zwischen der Beurteilung des Gesamteindrucks und der Beurteilung der einzelnen Qualitätsmerkmale nur geringe Korrelationen bestehen. Dies sollte Anlaß sein, die Praxis der visuellen Beurteilung zu hinterfragen.

b-Korrespondenzanalyse bipolarer Daten und nichtlineare Biplots

Mit Hilfe der Korrespondenzanalyse werden die Qualitätsbeurteilungen der Betriebe nun weiter verdichtet. Zunächst werden die absoluten Beiträge der Variablen zu den ersten beiden Dimensionen für die vier Sorte-Woche-Kombinationen miteinander verglichen und dann die zweidimensionalen Korrespondenzanalyseplots getrennt für die Sorten ‘Sierra’ und ‘Concerto’ sowie jeweils die Wochen 44 und 48 dargestellt und interpretiert. Es erfolgt abschließend eine zusammenfassende Beurteilung.

Nichtlineare Biplots, basierend auf einer Hauptkoordinatenanalyse, werden im Anschluß als Alternative zur Korrespondenzanalyse diskutiert.

- Abbildung A6 zeigt die über die positiven und negativen Pole akkumulierten, absoluten Beiträge der Variablen in den ersten beiden Dimensionen.

⇒ Auffällig ist, das nur ein einziges Merkmal, nämlich die Beurteilung der Wurzelqualität in allen vier Analysen einen wesentlichen Beitrag zur Gestaltung der Plots leistet, in drei von vier Fällen auf der zweiten Hauptachse. Die Beurteilung der Gesamtqualität beeinflußt die Darstellungen für die Sorte ‘Sierra’ stärker als für die Sorte ‘Concerto’. Allerdings wird auch sichtbar, daß zwischen den vier Abbildungen recht große Unterschiede bestehen, und somit offensichtlich die größten Quellen der Variabilität in allen vier Analysen bei anderen Merkmalen zu suchen sind, obwohl der Anteil der durch die ersten beiden Dimensionen ‘erklärten’ Inertia in allen Fällen annähernd gleich ist (um 60%).

- Abbildung A7 beinhaltet den Plot für ‘Sierra’ in Woche 44.

⇒ Die geringen Korrelationen zwischen den Variablen sind ebenso zu erkennen (approximiert durch die Winkel zwischen den Linien) wie die Unterschiede der Variabilität und der Polarisation des Mittels der einzelnen Variablen (approximiert durch die Länge der Linien und ihren Schnittpunkt mit dem Ursprung). Alle Variablen weisen eine gewisse Polarisation des Mittels auf, wobei der Ursprung in allen Fällen dem positiven Pol näher ist als dem negativen, was ein Hinweis auf insgesamt hohe mittlere Boniturwerte ist. Besonders stark ist die Polarisation des Mittels bei dem Kriterium Vergilbung (9,143), das heißt diese Variable ist am stärksten durch Bonituren nahe der Extreme der Bonitursкala geprägt. Die Objekte konzentrieren sich tatsächlich an einem der beiden Endpunkte der Bonitursкala (in diesem Fall am positiven Pol). Dies wird auch durch den sehr hohen Wert bei der Polarisation der Objekte von 10,667 bei dieser Variablen unterstrichen. Die Darstellungsqualität in zwei Dimensionen ist für alle Variablen mit Ausnahme des Knospenbesatzes recht gut (Qualität (und zwar Qualität im Sinne der Korrespondenzanalyse, siehe 2.1.3) in allen Fällen größer als 0,5, Knospenbesatz 0,1968). Auch die Darstellungsqualität der Objekte ist relativ hoch (2, 3, 8, 13, 18 und 19 über 0,7), allerdings gibt auch sehr schlecht repräsentierte Objekte, insbesondere die Betriebe 4 und 16, die in der Abbildung direkt am Ursprung liegen und ihre Lage in einer dritten Dimension bei einer Qualität in den ersten beiden Dimensionen von unter 0,1 haben. Auffällig ist die durch die

Korrespondenzanalyse bedingte starke Hervorhebung von Objekten mit extremen Werten, wie zum Beispiel des Betriebes 8 mit einer sehr geringen und der Betriebe 18 und 19 (liegen übereinander) mit einer sehr hohen Bewertung der Wurzelqualität. Tendenziell läßt sich festhalten, daß von oben nach unten die Boniturwerte mit Ausnahme der Wurzelbeurteilung, abnehmen und von links nach rechts die Wurzelqualität und (merkwürdigerweise) die Anfälligkeit für Krankheiten zunimmt.

– Abbildung A8 zeigt den Plot für ‘Sierra’ in Woche 48.

⇒ Erkennbar wird die deutliche Abnahme der Polarisation des Mittels, was in diesem Fall einer Abnahme der Boniturwerte entspricht, die Beeinflussung der ersten Dimension durch die Variablen Gesamtbeurteilung und Krankheiten, sowie die Bestimmung der zweiten Dimension durch die Merkmale Knospen- und Wurzelbeurteilung. Besonders auffällig ist die Position der Betriebe 3, 7 und 19 im Bereich der negativen Pole von Gesamt- und Krankheitsbeurteilung, von Betrieb 8 am negativen Pol der Knospen- und Wurzelbeurteilung und Betrieb 12 am positiven Pol der Gesamtbeurteilung. Die übrigen Betriebe trennen sich recht bemerkenswert in zwei Gruppen, eine Gruppe mit eher hohen Werten bei Gesamt-, Knospen- und Wurzelbeurteilung und niedrigen Werten bei den übrigen Beurteilungen (Betriebe 2, 3, 4, 5, 13, 15, 20) und eine zweite Gruppe mit hohen Werten bei Vergilbungs-, Krankheits- und Welkebeurteilungen und niedrigen Werten bei den anderen Merkmalen (Betriebe 9, 10, 11, 14, 16, 18).

– Abbildung A9 beinhaltet den Plot für ‘Concerto’ in Woche 44, Abbildung A10 für ‘Concerto’ in Woche 48.

⇒ Die Abnahme der Polarisation, und damit der Bonituren, tritt bei ‘Concerto’ wie bei ‘Sierra’ hervor. Der starke Einfluß der Wurzelbeurteilung in der zweiten Dimension, nahezu im rechten Winkel zu den übrigen Merkmalen in beiden Plots, deutet auf annähernde Unabhängigkeit zwischen der Beurteilung der Wurzelqualität und der Beurteilung der übrigen Merkmale hin. Diese scheinen ansonsten höher korreliert zu sein als bei ‘Sierra’, allerdings ist die Qualität der Darstellung, vor allem in A10 zum Teil sehr gering (Gesamt, Vergilbung, Welke und Krankheiten deutlich unter 0,5). Beide Plots weisen eine recht starke Gruppierung der Betriebe in der ersten Dimension auf, mit den Betrieben 1, 2, 5, 10, 13, 14, 15 und 18 an den positiven Polen und den Betrieben 3, 7, 8, 11, 19 an den negativen Polen in Woche 44, sowie den Betrieben 5, 13, 14, 15, 16, 18, 20 am negativen Pol der Knospenbeurteilung und an den positiven Polen der anderen Variablen (außer Wurzeln), sowie den Betrieben 1, 2, 3, 6, 7, 8, 10, 12 und 17 am positiven Pol der Knospenbeurteilung, aber an den negativen Polen der übrigen Variablen (außer Wurzeln) in Woche 48.

∗ Bei den zweidimensionalen Korrespondenzanalyseplots in den Abbildungen A7 bis A10 handelt es sich um jeweils zweidimensionale Approximationen an eine 6-dimensionale Wirklichkeit. Bei einer Interpretation derartiger Plots ist zu beachten, daß mit dieser Approximation ein gewisser

Informationsverlust einhergeht und Fehlrepräsentationen durchaus möglich sind, zumal wie in diesem Beispiel der Anteil der von den ersten beiden Dimensionen ‘erklärten’ Inertia in allen Abbildungen relativ gering ist. Sie liefern aber eine übersichtliche graphische Zusammenfassung einer Vielzahl von Informationen zu den jeweiligen Daten. Allerdings handelt es sich bei den Korrespondenzanalyseplots nicht um Biplots im eigentlichen Sinne, da die Distanz zwischen Objektkoordinaten und Variablenkoordinaten nicht definiert ist. Eine echte Biplotinterpretation im Sinne interpolativer Biplots ist für ordinalskalierte Variablen entweder durch die Konstruktion von Interpolationsregionen (siehe 2.2.2) oder die im folgenden Abschnitt behandelten nichtlinearen, interpolativen Biplots möglich.

– Die nichtlinearen Biplots der Abbildungen A12 bis A15 basieren auf einer Hauptkoordinatenanalye der am Mittelwert zentrierten Qualitätsbonituren. Das der Distanzmatrix der Hauptkoordinatenanalyse zugrunde liegende Maß ist die wegen ihrer sinnvollen ordinalen Interpretation gewählte City-Block-Distanz. Die Abbildungen A11a bis A11h zeigen die Konfigurationen der Betriebe bei beiden Sorten in beiden Wochen mit und ohne überlagerten Multiple Spanning Tree.

⇒ Die Eigenwerte der ersten beiden Dimensionen sind vergleichsweise gering, in allen Fällen ist der Anteil ‘erklärter’ Varianz kleiner als 50%. An der Überlagerung durch den Multiple Spanning Tree wird deutlich, daß es an verschiedenen Stellen zu recht erheblichen Verzerrungen in der dimensionserniedrigten Darstellung kommt (außer in A11f) und manche Objekte offensichtlich recht schlecht repräsentiert sind (zum Beispiel liegen in A11h die Betriebe 15 und 16 in der Abbildung sehr dicht beieinander, während die ihnen tatsächlich am nächsten liegenden Betriebe, das sind der Betrieb 20 (für 15) und der Betrieb 6 (für 16) in der zweidimensionalen Approximation sehr viel weiter entfernt sind). Die Lage der Betriebe im Koordinatensystem zueinander ist jedoch der vorangehenden Darstellung durch die Korrespondenzanalyse recht ähnlich (außer bei ‘Concerto’, Woche 48), was auch aufgrund der engen Beziehung zwischen Hauptkoordinatenanlyse und Korrespondenzanalyse nicht verwundert.

– Bei den nichtlinearen Biplotbahnen der Abbildungen A12 bis A15 handelt es sich der Übersichtlichkeit halber um an den Schnittpunkten der Achsen um den Faktor 6 (das entspricht der Anzahl der Variablen) gestreckte Achsen, so daß theoretisch eine Interpolation durch das einfache Auffinden des Zentroids der jeweils zutreffenden Variablenmarker möglich ist.

∗ Allerdings fällt auf, daß es in einigen Fällen nicht einmal für die Orginalobjekte eine hundertprozentig genaue Interpolation geben kann. Dies ist bedingt durch die Verwendung der nicht euklidisch einbettbaren City-Block-Distanz. Nur echte euklidische und euklidisch einbettbare Distanzmaße führen zu einer exakten Interpolation. Sowohl die nicht exakte Interpolation, als auch der zum Teil recht stark gekrümmte Verlauf der Biplotbahnen sind ein Indiz für das grundsätzliche Problem, das die Verwendung nicht-euklidischer Distanzmaße im euklidischen Referenzsystem mit sich bringt.

⇒ Inhaltlich wird die Interpretation durch die nichtlinearen Biplots nicht wesentlich verändert. Allerdings sind den Plots noch weitere Informationen zu entnehmen. Die Marker an den Endpunkten der Biplotbahnen bezeichnen Minimum und Maximum der jeweiligen Variablen und so ergibt sich gleichzeitig auch eine Information zu deren Spannweite. Der Schnittpunkt der Biplotbahnen liegt am - im Kontext ordinalskalierter Variablen weniger aussagekräftigen - Mittelwert der Variablen. Die wenigstens approximativ mögliche Interpolation soll nicht zu einer überzogenen numerischen Interpretation verwendet werden, zumal der Anteil 'erklärter' Varianz in allen Beispielen relativ gering ist; dennoch liefert die Nähe oder Ferne der Betriebe zu den jeweiligen Markern weitere Hinweise auf deren wahrscheinliche Werte.

c-Gemeinsame Betrachtung durch ordinale mehrdimensionale Skalierung der Korrelationsmatrix und generalisierte Prokrustes-Analyse

Die Qualitätsbeurteilungen, die an zwei Zeitpunkten und an zwei Sorten bei sechs Merkmalen, also in insgesamt vier Kombinationen vorgenommen wurden, sollen nun einer gemeinsamen Betrachtung unterzogen werden.

- Aufbauend auf der Spearman-Korrelationsmatrix, die die Korrelationen für die Sorte-Merkmal-Kombinationen, dem Multi-Trait-Multi-Method-(MTMM)-Ansatz[37] folgend, beinhaltet (siehe Übersicht A8), wird für beide Beurteilungswochen getrennt zunächst eine Hauptkoordinatenanalyse gerechnet (siehe Übersicht A9), und die aus ihr gewonnene Distanzmatrix dann einer ordinalen mehrdimensionalen Skalierung unterzogen. Der stress-Wert für Woche 44 ist mit 0,1073 recht gut, und selbst der stress-Wert für Woche 48 ist mit 0,1774 noch akzeptabel. Die Plots sind in den Abbildungen A16a bis A16d zusammengefaßt. Die Abbildungen A16a und A16c verwenden zur Kennzeichnung der Punkte die Bezeichnung der Sorte (Si für 'Sierra', Co für 'Concerto'), die Abbildungen A16b und A16d die Bezeichnug der Merkmale (ges für Gesamt, kno für Knospenbesatz, wur für Wurzelqualität, gil für Vergilbung, wel für Welke, kra für Krankheiten). A16a und A16b beziehungsweise A16c und A16d beinhalten also die gleichen Konfigurationen, nur mit unterschiedlicher Kennzeichnung der Punkte.

⇒ Die Abbildungen A16a und A16b weisen eine deutliche Konzentration der in Woche 44 bestimmten Merkmale für 'Sierra' im unteren Teil und für 'Concerto' im oberen Teil der Abbildung auf, was auf eine geringe Korrelation der Boniturwerte zwischen den Sorten und eine stärkere Korrelation der Qualitätsmerkmale innerhalb der Sorten hindeutet. Dies trifft insbesonders für 'Concerto' zu, wo die Merkmale Gesamteindruck, Knospenbesatz, Vergilbung, Welke und Krankheiten dicht beieinander liegen und somit positiv miteinander korreliert sind, während sich die Wurzelqualität von dieser Variablengruppe deutlich absetzt, was auf eine

[37] Siehe BORG & GROENEN, 1997.

negative Korrelation zu den übrigen Merkmalen schließen läßt. Eine aus kulturtechnischer Sicht nur schwer erklärbare Beobachtung. Dieser Plot verstärkt damit die schon in Abbildung A9 gemachte Beobachtung. Bei ‘Sierra’ nimmt das Merkmal Gesamtbeurteilung eine zentrale Stellung ein. Der Gesamteindruck ist mit allen übrigen Merkmalen bei ‘Sierra’ positiv korreliert. Knospenbesatz und Wurzelbesatz auf der einen Seite, und Vergilbung, Welke und Krankheiten auf der anderen Seite bilden zwei Variablengruppen, deren Variablen innerhalb positiv, aber mit den Variablen der anderen Gruppe negativ korrelieren. Schließlich liegt der Gesamteindruck beider Sorten im Plot der mehrdimensionalen Skalierung relativ nahe beieinander, so das es sich bei diesem Merkmal um das zwischen den Sorten offensichtlich am stärksten korrelierte Merkmal handelt.

⇒ Die Abbildungen A16c und A16d zeigen die Plots für Woche 48, die sich gegenüber Woche 44 recht stark verändert haben. Die Korrelationen innerhalb der Sorten haben durchweg abgenommen, die Sortenpositionen sind wesentlich weniger kompakt, während die paarweisen Merkmalskombinationen in fast allen Fällen zugenommen haben. Dies trifft vor allem auf die nun sehr nahe beieinanderliegenden Merkmale Gesamteindruck, Wurzelqualität und Welke zu. Bei einem allgemeinen Rückgang der Qualität von Woche 44 zu Woche 48 (siehe A2) und einer Zunahme der Variabilität der Beurteilungen innerhalb jedes Betriebes vor allem bei ‘Concerto’ (siehe A3), sind die Bonituren in Woche 48 über beide Sorten offensichtlich etwas harmonischer und einheitlicher als in Woche 44. Während in Woche 44 ‘Sierra’ noch eine deutlich von ‘Concerto’ abweichende Beurteilungsstruktur aufweist, wird sie ihr in Woche 48 sehr viel ähnlicher. Dies mag Anlaß zu der Vermutung sein, daß in der Zeit nach dem Produktionsende der Einfluß der einzelnen Betriebe stärker zutage getreten ist. Während anfangs Sortenunterschiede dominieren, sind es später die besseren oder schlechteren Bewertungen der Qualitäten der einzelnen Betriebe, was als Indiz für die Wirkung einer inneren Qualität herangezogen werden könnte. Allerdings mag die stärkere Beziehung der Bonituren desselben Merkemals der beiden Sorten in Woche 48 auch zumindest teilweise durch einen Zugewinn an Beurteilungsroutine durch die bewertende Person zu erklären sein.

∗ Die Darstellung der Korrelationsmatrix in der Form unterschiedlich gekennzeichneter Plots nach ordinaler mehrdimensionaler Skalierung erweist sich als ein sehr hilfreiches Vorgehen, um Beziehungszusammenhänge aufzuspüren und graphisch abzubilden. Beim Vergleich mit den Korrelationsmatrizen (Übersicht A8) fällt auf, daß tatsächlich eine vernünftige Abbildung wesentlicher Zusammenhänge zustande gekommen ist. Allerdings darf nicht übersehen werden, daß die absolute Größe der Korrelationskoeffizienten gering ist, und daher auch in diesem Beispiel vor einer Überinterpretation gewarnt werden muß.

– In der folgenden Prokrustes-Analyse geht es um die Beurteilung der relativen Lage der Betriebe zueinander. Abbildung A2 ist bereits zu entnehmen, daß es zu einem Qualitätsrückgang zwischen Woche 44 und 48 gekommen ist. Die Abbildungen A4 und A5 verdeutlichen, daß es

sich bei ‘Sierra’ bei den dargestellten Merkmalen um die insgesamt etwas besser beurteilte Sorte handelt. Das heißt also, daß zwischen den Mittelwertsvektoren der Qualitätsbonituren von ‘Sierra’ Woche 44, ‘Sierra’ Woche 48, ‘Concerto’ Woche 44 und ‘Concerto’ Woche 48 erkennbare Unterschiede bestehen.

⇒ Werden die Boniturdaten direkt mit der Prokrustes-Analyse analysiert, so bestätigt der hohe Varianzwert für die Streuung zwischen den vier Konfigurationen von 1064,9 (initial between-configurations s.s.), das sind 40,4% der Gesamtstreuung, diese Unterschiede. In der Prokrustes-Analyse erfolgt als erster Schritt die Verschiebung der Konfigurationen auf einen gemeinsamen Ursprung, so daß diese Mittelwertsunterschiede eliminiert werden, da nicht diese im Mittelpunkt der Analyse stehen, sondern die relative Lage der Objekte zueinander.

– Es werden nun nicht die Orginalboniturwerte, sondern die durch die oben (b-) beschriebene Korrespondenzanalyse ermittelten Objektkonfigurationen der Prokrustes-Analyse unterzogen. Abbildung A17 faßt diese Konfigurationen noch einmal für die vier Sorte-Woche-Kombinationen in den ersten beiden Dimensionen zusammen. Nach erfolgter Prokrustes-Analyse ergeben sich Skalierungsfaktoren und Rotationsmatrizen, die die Orginalkonfigurationen (in allen Dimensionen) so verändern, daß sie zu einer größtmöglichen Deckung mit der Konsens-Konfiguration gelangen. Die durch die Skalierungsfaktoren und Rotationsmatrizen veränderten Orginalkonfigurationen sind in Abbildung A18 (für ‘Sierra’) und A19 (für ‘Concerto’) beziehungsweise in Abbildung A21 (für Woche 44) und A22 (für Woche 48) in den ersten beiden Dimensionen abgebildet. Sie ermöglichen eine bessere Vergleichbarkeit (bei gleicher Darstellungsgüte) als die in A17 abgebildeten Orginalkonfigurationen. Die jeweilige Konsens-Konfiguration ist in Abbildung A20 (getrennt für ‘Sierra’ und ‘Concerto’) beziehungsweise in Abbildung A23 (getrennt für Woche 44 und Woche 48) und in Abbildung A24 (für alle Kombinationen) zu finden. Der Anteil ‘erklärter’ Varianz durch die ersten beiden Dimensionen der Konsens-Konfigurationen liegt bei 60%.

⇒ Bei beiden Sorten kommt es zwischen den beiden Wochen zu einer recht guten Übereinstimmung zwischen den Konfigurationen. Das Residuum liegt bei ‘Sierra’ bei 0,536 (das entspricht 26,8%) und bei ‘Concerto’ sogar nur bei 0,441 (das entspricht 22,1%). Diese recht gute Übereinstimmung wird in A18 und A19 durch die Nähe der dunkelblauen (Woche 44) und der hellblauen (Woche 48) Punkte mit gleicher Bezeichnung verdeutlicht, zum Beispiel in A19 die Positionen der Betriebe 7, 8, 11, 13, 14, 15 und 18. Das Maß der Übereinstimmung läßt den Schluß zu, daß die Ähnlichkeitsbeziehungen der Betriebe innerhalb der Sorten zwischen den Wochen relativ stabil sind, bei ‘Concerto’ noch in höherem Maß als bei ‘Sierra’. Ähnlich verhält es sich bei der Betrachtung der beiden Beurteilungswochen. In Woche 44 tritt bei einem Residuum von 0,529 (das entspricht 26,5%), die nahezu identische Lage der Betriebe 2, 3, 13, 14, und 16 hervor (Abbildung A21). Demgegenüber ist das Residuum in Woche 48 mit 0,589 (das entspricht 29,5%) etwas höher, das heißt, die relative Lage der Betriebe weicht in Woche

48 (siehe Abbildung A22) in den sortenbezogenen Konfigurationen stärker voneinander ab als in Woche 44. Zusammmenfassend darf also festgestellt werden, daß die Konfigurationen von ‘Concerto’ stabiler sind als die von ‘Sierra’, und daß die Konfigurationen in Woche 44 stabiler sind als die in Woche 48. Einzelne Betriebe, die in den Konsens-Konfigurationen durch sehr große Residuen auffallen, die also ihre Lage in den verschiedenen Konfigurationen besonders stark verändern, sind (in Klammern die Residuen, also die individuellen Differenzen): Betriebe 7 (0,048), 12 (0,045) und 19 (0,046) bei ‘Sierra’; Betriebe 1 (0,045), 10 (0,056) und 19 (0,037) bei ‘Concerto’; Betriebe 2 (0,060), 7 und 11 (beide 0,047) und 17 (0,051) bei Woche 44; sowie Betriebe 3 (0,058), 12 (0,075) und 19 (0,052) bei Woche 48. Unterstrichen werden diese hohen Residuen durch die Dotplots in den Abbildungen 24a-d. So wird in 24a deutlich, daß sich bei ‘Sierra’ die Betriebe 7 und 19 im Mittel um 3 Einheiten von Woche 44 zu Woche 48 verschlechtert haben, während Betrieb 12 als einziger Betrieb von Woche 44 zu Woche 48 in der Beurteilung im Mittel nicht nach unten tendiert. Abbildung 24d zeigt demgegenüber, daß ‘Concerto’ bei Betrieb 3 in Woche 48 im Mittel als einziger Betrieb besser, ‘Concerto’ bei Betrieb 12 dagegen erheblich schlechter beurteilt wird als bei ‘Sierra’.

⇒ Beim Vergleich aller vier Konfigurationen (siehe Abbildung A25) wird deutlich, daß sich eine Vielzahl der Punkte am Ursprung der Abbildung konzentrieren. Dies ist ein Anzeichen dafür, daß es sich um relativ heterogene Konfigurationen handelt. Tatsächlich liegt das Residuum bei 1,659 (das entspricht 41,5%) und ist also deutlich höher als bei der paarweisen Prokrustes-Analyse. Es ist demnach sehr schwierig alle vier Kombinationen auf eine gemeinsame Konfiguration zu vereinen. Die relativen Ähnlichkeitsbeziehungen der Betriebe variieren zu stark von Sorte zu Sorte und von Woche zu Woche. Die einzelnen Beurteilungszeitpunkte stellen also Momentaufnahmen dar, aus denen sich sich nur schwerlich eine allgemeingültige Beziehung der Qualität der Betriebe untereinander ableiten läßt.

∗ Eine alternative Betrachtung derselben Daten kann durch eine gewichtete, ordinale mehrsdimensionale Skalierung erfolgen. Die Ergebnisse werden hier nicht gezeigt. Die stress-Werte in zwei Dimensionen sind sehr hoch, und die gemeinsame Konfiguration aller vier Konfigurationen ähnlich unscharf wie die der Prokrustes-Analyse.

d-Hierarchische Clusteranalyse

Da die bislang gemachten Beobachtungen nicht darauf hindeuten, daß es bezogen auf alle Merkmale, Sorten und Beurteilungszeitpunkte eine durchgängige Struktur bei den Betrieben gibt, die auf eine deutliche Gruppenbildung schließen ließe, soll die Clusteranalyse hier nur als ergänzendes, deskriptives Instrument eingesetzt werden.

– Abbildung A26 zeigt die Dendrogramme unterschiedlicher Clusteralgorithmen bei Verrechnung des gesamten Datensatzes, also der 24 Variablen der Wochen 44 und 48 bei ‘Sierra’ und ‘Concerto’.

⇒ Die Clusterverfahren produzieren, aufbauend auf einer durch die City-Block-Distanz gebildeten Proximitätsmatrix, naturgemäß unterschiedliche Gruppenbildungen. Allerdings treten immer wieder die Betriebe 3, 8, 7, 11 und 19 durch eine etwas besondere Stellung hervor, die sich von einem recht kompakten Bereich der übrigen Betriebe mehr oder weniger stark abheben. Die Orginaldaten zeigen, daß der Betrieb 3 dabei durchgängig mit Ausnahme bei 'Concerto' in Woche 48 sehr schlecht beurteilt wird, der Betrieb 8 in der Regel schlechte Beurteilungen bei den Variablen Gesamteindruck, Knospenbesatz und Wurzelqualität und durchgängig recht gute Beurteilungen bei den Variablen Vergilbung, Welke und Krankheiten erhält. Bei der Gruppe der Betriebe 7, 11 und 19 verhält es sich annähernd umgekehrt, während schließlich die große Gruppe der restlichen Betriebe eine große Streuung der Boniturwerte ohne eine klar erkennbare Struktur aufweist.

3.1.2.2 Variablenset 2 - Analyse der Kultursubstrate

a-Einführende Datenanalyse

– Die Betrachtung der Substratanalysewerte wird mit der Darstellung der 12 Variablen in Form einer Scatterplotmatrix in Abbildung A27 begonnen. Die Betriebe, die in der vorangegangenen Clusteranalyse der Qualitätsmerkmale eine besondere Position eingenommen haben, sind farblich hervorgehoben, und zwar der Betrieb 3 rot, der Betrieb 8 blau, die Betriebe 7, 11 und 19 grün. Die übrigen Betriebe sind schwarz gekennzeichnet.

⇒ Die Scatterplotmatrix zeigt auffällige Korrelationen zwischen N, K und Salz vor allem in Woche 23. In den anderen beiden Wochen sind die Beziehungen bei weitem nicht so stark ausgeprägt. Auch liegen offensichtlich zwischen den Terminen keine starken Korrelationen vor. Dies bestätigt die Tatsache, daß sich selbst bei zeitlich nahe beieinander liegenden Untersuchungen aufgrund der vielfältigen Einflußmöglichkeiten der Kultivateure stark voneinander abweichende Ergebnisse ergeben können. Insbesondere in Woche 41 werden die N, K und Salz-Plots durch einzelne sehr extreme Werte so beeinflußt, daß kaum noch etwas von der Beziehung der Merkmale zueinander im Scatterplot erkennbar ist. Interessanterweise handelt es sich bei den auffälligen Werten um die Betriebe 3 (N41), 3 und 8 (K41), und 3 und 19 (SALZ41), also um Betriebe, die auch schon bei der Qualitätsbeurteilung aufgefallen sind. Auch in den übrigen Wochen liegen die Betriebe der drei kleinen Cluster häufig am oberen Ende der Skala, so daß es sich demnach um Betriebe handelt, die recht stark gedüngt haben. Die Rangkorrelelationsmatrix liefert in Übersicht A10 die numerische Information zu den Korrelationen (ohne Betrieb 5, siehe folgender Abschnitt).

– Die Substratanalysen konnten in Woche 41 für den Betrieb 5 nicht durchgeführt werden. Es tauchen daher im Datensatz insgesamt vier fehlende Werte auf. Vor einer Schätzung dieser fehlenden Werte durch das Verfahren nach BEALE & LITTLE, 1975, soll der Datensatz auf Multi-normalverteilung überprüft werden, da das genannte Einsetzungsverfahren diese

voraussetzt.

⇒ Übersicht A11 zeigt, daß Tests sowohl auf univariate (marginal) als auch auf bivariate (bivariate angle) und multivariate (radius) Normalverteilung vielfach signifikante Ergebnisse erbringen, daß es also deutliche Hinweise auf ein Abweichen von der Normalverteilung gibt, obwohl die Testergebnisse zum Teil erheblich voneinnader abweichen. Der Versuch einer Schätzung der fehlenden Werte nach BEALE & LITTLE, 1975, produziert demzufolge nur bedingt brauchbare, für N41 sogar negative Schätzer (N41 geschätzt -265,7; K41 geschätzt 144,8; SALZ41 geschätzt 0,048; PH41 geschätzt 5,317). Es wird daher hier so weiter vorgegangen, daß die Auswertung der Substratanalysewerte ohne den Betrieb 5 durchgeführt wird.

– Weiterhin soll überprüft werden, ob es Anzeichen für multivariate Ausreißer gibt. Hierzu wird das Verfahren nach CAMPBELL, 1980, eingesetzt.

⇒ Obwohl einige Betriebe recht hohe Mahalanobis-Distanzen besitzen, insbesondere die Betriebe 3 und 19, ist das Indiz für echte Ausreißer gering. Es wird daher in der folgenden Hauptkomponentenanalyse mit der normalen Kovarianzmatrix der standardisierten Werte gerechnet und nicht mit einem robusten Schätzer der Korrelationsmatrix. Die Standardisierung ist erforderlich, da die einzelnen Variablen sehr unterschiedliche Maßeinheiten besitzen und stark voneinander abweichende Varianzen aufweisen (siehe Übersicht A2).

b-Hauptkomponentenanalyse I - Anzahl der 'wesentlichen' Dimensionen

Bevor die Substratanalysewerte in Form von Hauptkomponentenanalyse-Biplots graphisch dargestellt werden, soll zunächst die Frage untersucht werden, wieviele Komponenten eigentlich notwendig sind, um die in den Daten enthaltenen Informationen mit möglichst geringem Informationsverlust abzubilden. Es werden diskutiert: Screeplot, CUSUM-Diagramm und Anteil 'erklärter Varianz'; Velicers partielle Korrelations-Methode; die PRESS-Statistik; sowie Residuenplots und -tests. Auf die Anwendung von Signifikanztests zur Bestimmung der Anzahl 'wesentlicher' Dimensionen wird sowohl aufgrund des geringen Stichprobenumfangs, als auch aufgrund der Verwendung der standardisierten Werte (und damit der Korrelationsmatrix) verzichtet.

– Übersicht A13 zeigt den Screeplot nach erfolgter Hauptkomponentenanalyse der standardisierten Substratanalysewerte (ohne Betrieb 5). Zusätzlich sind die Eigenwerte, sowie der Anteil 'erklärter' Varianz jeder Komponente, und der akkumulierte Anteil 'erklärter' Varianz wiedergegeben. Zusätzlich informieren die Del1, Del2 und Del3-Spalten über die ersten Differenzen (das heißt Del1 ist die Differenz zweier per-1000-Werte (also zum Beispiel 416 - 183 = 234 (Rundungsfehler)), Del2 die Differenz zweier Del1-Werte (also zum Beispiel 234 - 45 = 189) und Del3 die Differenz zweier Del2-Werte (also zum Beispiel 189 - 8 = 181)). Diese Differenzen geben Hinweise auf den Verlauf des Screeplots und eventuelle Plateaus (angezeigt durch hohe Del2 und Del3-Werte).

⇒ Es wird deutlich, daß die erste Hauptkomponente eine herausgehobene Stellung einnimmt. Die

Komponenten zwei bis vier liegen erheblich darunter und es liegt durchaus nahe zwischen der ersten und den übrigen Hauptkomponenten eine Art ‘Bruchstelle’ zu sehen. Die hohen Del2 und Del3 Werte der zweiten und dritten Komponente deuten darüberhinaus an dieser Stelle auf eine deutliche, auch in der Graphik sichtbare Abflachung hin. Der absolute Anteil ‘erklärter’ Varianz durch die erste Hauptkomponente ist mit 42% aber relativ gering. Erst mit sechs Komponenten gelingt eine Abdeckung von über 90% der Varianz. Nach dem sogenannten Kaiser-Kriterium (Auswahl aller Hauptkomponenten mit Eigenwerten größer 1) sind in diesem Beispiel vier Hauptkomponenten zu betrachten. Drei unterschiedliche Kriterien liefern somit drei unterschiedliche Ergebnisse.

– Eine Bereicherung des Screeplots stellt die Darstellung als CUSUM-Diagramm dar. Der Eigenwert jeder Hauptkomponente ist unterteilt in die Anteile der einzelnen Variablen, die diese zum jeweiligen Eigenwert beitragen. Es liefert somit nicht nur eine Information über die relative Bedeutung jeder Komponente, sondern auch über die Variablen, die im wesentlichen diese Komponente bestimmen. Abbildung A28 beinhaltet das CUSUM-Diagramm für die Substratanalysewerte.

⇒ Neben der großen Bedeutung der ersten Hauptkomponente wird nun sichtbar, daß diese vor allem durch die Werte aus Woche 23 und 29 bestimmt ist, und hier vor allem durch N-, K- und Salz-Messungen. Während die Variablen aus Woche 23 fast vollständig in der ersten Dimension abgebildet sind, verteilt sich die Bedeutung der Variablen aus Woche 41 fast zu gleichen Teilen auf die ersten drei Komponenten. Die pH-Wert-Messungen sind in der ersten Dimension sehr schlecht dargestellt, nehmen dafür aber in der zweiten zusammen mit der Salz-Messung, sowie auch in der dritten und vierten Dimension eine wichtige Stellung ein. Die erste Dimension bildet also vor allem N-, K- und Salzgehaltsunterschiede in den Wochen 23 und 29 zwischen den Betrieben ab, während die zweite Dimension die Betriebe nach pH-Wert und Salzgehaltsunterschieden differenziert.

∗ Zwischen CUSUM-Diagramm und der oben (siehe Abbildung A6) verwendeten Darstellung der absoluten Beiträge der Variablen besteht eine enge Beziehung. Es handelt sich nämlich bei den absoluten Beiträgen lediglich um den prozentualen Anteil der einzelnen Variablen an dem jeweiligen Eigenwert jeder Komponente.

– Die partielle Korrelationsprozedur nach VELICER, 1976, verwendet als als Entscheidungskriterium die f_q-Werte. Es ist die Anzahl von Komponenten ausreichend, bei denen f_q sein Minimum hat. Die PRESS-Statistik nach EASTMENT & KRZANOWSKI, 1982, führt zur Berechnung der W-Werte, und die Anzahl ‘wesentlicher’ Komponenten wird durch W-Werte von größer 1 charakterisiert. Abbildung A29a zeigt einen Plot der f_q-Werte, Abbildung A29b einen Plot der W-Werte. Die Übersichten A14a und A14b geben die entsprechenden numerischen Ergebnisse.

⇒ In beiden Fällen deuten die Verfahren darauf hin, daß die erste Hauptkomponente zu einer angemessenen Abbildung der Daten ausreicht. Es bestätigt sich damit die bekannte Beobachtung, daß diese Verfahren weniger Komponenten auswählen als das Kaiser-Kriterium oder das „Anteil 'erklärter' Varianz von über 90%"-Kriterium.

– Die Abbildung A30 zeigt einen Dotplot der Hauptkomponenten-Residuen bei Verwendung von einer (Kreis) beziehungsweise von zwei (Kreuz) Hauptkomponenten. Die Übersichten A15a und A15b liefern die Werte der Residuen und die kritischen Werte bei einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$.

⇒ In beiden Fällen sind die Residuen keines der betrachteten Objekte größer als der errechnete kritische Wert. Allerdings haben die Betriebe 3, 9, 12 und 19 (bei einer Hauptkomponente) sowie die Betriebe 3 und 12 (bei zwei Hauptkomponenten) recht hohe Residuen, so daß zumindest diese Objekte in der ein- beziehungsweise zweidimensionalen Darstellung nicht sehr gut repräsentiert sind. In einigen Fällen führt die Hinzunahme der zweiten Komponente zu einer deutlichen Verringerung der Residuen (zum Beispiel beim Betrieb 9) in anderen Fällen bleibt das Residuum nahezu unverändert (zum Beispiel bei Betrieb 14). Aufgrund der geringen Stichprobengröße ist die Aussagekraft der kritischen Werte schwach. Einen deutlichen Hinweis auf die Notwendigkeit der Betrachtung von mehr als zwei Dimensionen liefert in diesem Beispiel jedoch auch die Residuenanalyse nicht.

c-Hauptkomponentenanalyse II - Hauptkomponenten-Biplots

Es folgt nun die Darstellung der Substratanalysewerte in Form der von GOWER & HAND, 1996, beschriebenen Hauptkomponenten-Biplots.

– Die Biplots mit den Interpolationsmarkern in den Abbildungen A31, A32 und A33 zeigen die Konfigurationen der Betriebe und, der Übersichtlichkeit halber, die Variablenachsen in Woche 23 (A31), in Woche 29 (A32) und in Woche 41 (A33).

⇒ Da es sich um standardisierte Werte handelt, ist die Möglichkeit der graphischen Interpolation an der Abbildung natürlich eingeschränkt, da nicht die Originalwerte, sondern die standardisierten Werte an den Biplotachsen angezeigt werden. Auffällig sind vor allem die Betriebe 3 und 14 mit sehr hohen K-Werten an allen Meßzeitpunkten. Daneben fällt eine Gruppe von Betrieben mit hohen pH-Werten in Woche 23 und niedrigen pH-Werten in den Wochen 29 und 41 auf (Betriebe 2, 7, 8, 9, 13), die im Bezug auf die übrigen Merkmale recht stark streut. Eine zweite Gruppe, die alle restlichen Betriebe beinhaltet, liegt im Bereich mittlerer und unterdurchschnittlicher pH-Werte zu Kulturbeginn (Woche 23), und mittleren und überdurchschnittlichen pH-Werten in den Wochen 29 und 41. Bei N-, K- und Salz-Messungen liegen diese Betriebe im mittleren oder unterdurchschnittlichen Bereich.

– Die Prediktionsmarker sind in den Abbildungen A34 (Woche 23), A35 (Woche 29) und A36 (Woche 41) wiedergegeben. Sie bieten die Möglichkeit zur graphischen Abschätzung der Werte

der einzelnen Variablen bei den Betrieben. Ohne eine weitere Verrechnung ist die graphische Prediktion aber aufgrund der Verwendung der strandardisierten Werte auch in diesem Fall nicht leicht interpretierbar. Im für die Hauptkomponenten-Biplots entwickelten Genstat-Code wird jedoch interaktiv mit Hilfe des DREAD Befehls zunächst der standardisierte Wert für die Prediktion durch eine rechtwinklige Projektion vom Betriebspunkt auf die Biplotachse verwendet, der dann intern auf den Orginalwert umgerechnet wird. Im Ausdruck erscheint dann die Biplot-Approximation des jeweiligen Variablenwertes des ausgewählten Betriebes.

⇒ Beispielhaft zeigt die Übersicht A16 die Ergebnisse der interaktiven Prediktion für den Betrieb 3 beim Salzgehalt in Woche 23, für den Betrieb 2 beim K-Wert in Woche 29, und für den Betrieb 19 beim N-Wert in Woche 41. Je nach Darstellungsgüte der Variablen (gemessen als Quadratsumme der Koeffizienten der Eigenvektoren der betreffenden Variablen der betrachteten Dimensionen) und der Betriebe wird eine mehr oder weniger gute Prediktion erreicht.

∗ Das interaktive Vorgehen erlaubt beim Prediktions-Biplot die Extraktion einer Vielzahl von Informationen zu den Betrieben bei der Betrachtung einer einzelnen Graphik und stellt somit ein wichtiges Intrument in der Kommunikation der Ergebnisse der Substratanalysewerte dar, das so durch einen herkömmlichen Biplot, wie er in Abbildung A37 wiedergegeben ist, bei weitem nicht erreicht werden kann.

3.1.2.3 Variablenset 3 - Aufzeichnung der Kulturmaßnahmen

Der dritte Variablensatz beinhaltet eine Zusammenstellung mehrerer Variablen, die sich auf die Kulturführung in den einzelnen Betrieben beziehen. Zum Teil werden direkt bestimmte Merkmale verwendet (zum Beispiel Endstand in Pflanzen je m^2), zum Teil aus den Kulturaufzeichnungen abgeleitete Werte (zum Beispiel Verhältnis Pflanzen zu Kulturbeginn je m^2 zu Pflanzen im Endstand je m^2). Darüber hinaus wird dieses Variablenset in drei Untergruppen gegliedert: erstens Einstellung der Schattiersollwerte (Übersicht A3a), zweitens Platzbedarf und Rücken (Übersicht A3b) und drittens Verlauf der Temperaturführung (Übersicht A3c).

a-Hauptkomponentenanalyse der Schattiersollwerte

– Die Betriebe 11 und 18 operieren nicht mit Schattiersollwerten, so daß sie in dieser Auswertung nicht berücksichtigt werden können. Die Übersicht A3a zeigt bereits, daß im Mittel die Schattiersollwerte mit Zunahme der Kulturdauer zunehmen, das heißt, daß im Mittel zu Kulturbeginnn stärker schattiert, also dunkler kultiviert wird als im weiteren Kulturverlauf. Um zu einer Differenzierung zwischen den Betrieben zu gelangen, bietet sich die Hauptkomponentenanalyse der Kovarianzmatrix an. In diesem Fall ist eine Standardisierung der Werte nicht erforderlich. Die Ergebnisse der Hauptkomponentenanalyse sind in Übersicht A17 zusammengefaßt, während die Abbildungen A38a und A38b die Hauptkomponentenwerte der ersten und der zweiten Dimension in Form von Dotplots darstellen.

⇒ Die erste Hauptkomponente spielt in dieser Auswertung eine überragende Rolle. Sie repräsentiert 87,3% der Gesamtstreuung, die zweite Hauptkomponente repräsentiert nur noch weitere 8%. Damit decken die ersten beiden Hauptkomponenten bereits über 95% der Gesamtstreuung von 13 Ausgangsvariablen ab und stellen somit tatsächlich eine erhebliche Möglichkeit der Dimensionserniedrigung dar. Darüberhinhaus lassen sie eine schlüssige inhaltliche Interpretation zu. Die erste Komponente sortiert die Betriebe nach der Stärke der Schattierung insgesamt, das heißt Betriebe mit kleinen Hauptkomponentenwerten haben hohe Schattiersollwerte und somit insgesamt heller kultiviert und Betriebe mit hohen Hauptkomponentenwerten haben niedrige Schattiersollwerte, das heißt sie haben insgesamt dunkler kultiviert. Im Dotplot in Abbildung A38a stehen demnach die heller kultivierenden Betrieb oben und die dunkler kultivierenden Betriebe unten in der Abbildung. Auffällig ist, daß sich unter den fünf Betrieben mit der dunkelsten Kulturführung (7, 3, 19, 8 und 14), vier befinden (3, 7, 8 und 19), die auch schon in der der Qualitätsbeurteilung negativ aufgefallen sind. Allerdings liegt der Betrieb 14, der insgesamt am stärksten schattiert hat, bei der Qualitätsbeurteilung im mittleren bis oberen Bereich. Dennoch scheint diese Beobachtung eher auf ein Risiko durch zu dunkle als durch zu helle Kulturführung hinzudeuten (im Rahmen der in dieser Untersuchung eingestellten Schattiersollwerte). Die zweite Hauptkomponente, deren Hauptkomponentenwerte in Abbildung A38b wiedergegeben sind, trennt im wesentlichen nach der Veränderung der Schattiersollwerte im Kulturverlauf. So korrespondieren kleine Hauptkomponentenwerte mit hohen Schattiersollwerten, also heller Kultur, im zweiten Kulturabschnitt (ab Woche 29) und im Kulturverlauf ansteigenden Sollwerten, während hohe Hauptkomponentenwerte gleichbleibenden oder abnehmenden (dies trifft aber nur für Betrieb 8 zu) Schattiersollwerten entsprechen.

b-Ordinale mehrdimensionale Skalierung aller Merkmale

In der Folge werden nun alle Variablen des Variablenset 3 mit Hilfe der ordinalen mehrdimensionalen Skalierung verrechnet. Anstelle der ursprünglichen Schattiersollwerte werden nur die erste und die zweite Hauptkomponente verwendet. Da die Variablen sehr unterschiedliche Skalenniveaus besitzen, wird eine Proximitätsmatrix mit Hilfe des allgemeinen Ähnlichkeitskoeffizienten (nach GOWER & LEGENDRE, 1986) gebildet. Damit ergibt sich auch die Möglichkeit der Einbeziehung der Objekte mit fehlenden Werten. Durch Beziehung aller quantitativen Variablen auf ihre Spannweite, erfolgt die Gleichgewichtung aller Variablen.

- Übersicht A 18 zeigt die einzelnen Variablen und das jeweils verwendete Proximitätsmaß. Bei der ordinalen mehrdimensionalen Skalierung wird das least squares stress-Kriterium verwendet, es wird die Hauptkoordinatenanlysenkonfiguration als Ausgangskonfiguration eingesetzt, es werden dann 50 weitere, zufällige Konfigurationen gebildet und es erfolgt die primäre Behandlung gleicher Werte, das heißt es werden keine Begrenzungen auf Objekte mit identischen Werten in der Proximitätsmatrix gesetzt. Die Skalierung mit dem geringsten stress-

Wert (von 50 Ausgangskonfigurationen) wird weiter betrachtet. Übersicht A19 beinhaltet die Eigenwerte der Hauptkoordinatenanalyse, sowie die stress-Werte bei einer Skalierung in zwei, drei und vier Dimensionen beziehungsweise die Koordinaten der vierdimensionalen Lösung. Die Abbildungen A39 und A40 zeigen die Shepard-Plots für die drei Lösungen der mehrdimensionalen Skalierung (die grüne Line entspricht der monotonen Regression zwischen Dissimilaritäten und Distanzen, also den Disparitäten), die Abbildung A41 die Konfigurationen der Hauptkoordinatenanlyse und der mehrdimensionalen Skalierung in zwei Dimensionen.

⇒ In der vierdimensionalen Lösung wird ein recht niedriger stress-Wert von 0.0851 erreicht. Der Shepard-Plot in A41 bestätigt die gute Anpassung in vier Dimensionen und die deutliche Verbesserung gegenüber der Betrachtung von nur zwei oder drei Dimensionen. Die Darstellung in zwei Dimensionen führt zu einer erheblichen Verzerrung, wie aus den Graphiken in A41 durch die überlagerten Multiple Spanning Trees sichtbar wird; die mehrdimensionale Skalierung erbringt eine noch ein wenig bessere Darstellung als die Hauptkoordinatenanalyse.

– Es wird nun der Versuch unternommen, mehr als zwei Dimensionen abzubilden. Abbildung A42 enthält eine dreidimensionale Darstellung, A43 Andrews-Kurven der ersten vier Dimensionen, A44 einen Parallelkoordinatenplot der ersten vier Dimensionen, und Abbildung A45 schließlich ein Trellis-Display der dritten und vierten Dimension, konditioniert nach Werten der ersten und zweiten Dimension.

⇒ Die dreidimensionale Abbildung (A42), die auch noch rotiert werden könnte, verdeutlicht, wie schwierig es ist, Ähnlichkeitsbeziehungen in drei Dimensionen abzulesen, die zu einer Gruppierung der Betriebe führen könnte. Die Andrews-Kurven (A43) der ersten vier Dimensionen lassen eine solche Gruppierung allerdings vermuten, und mögliche Gruppen sind durch unterschiedliche Farben gekenneichnet (jede Kurve ist mit der zugehörigen Betriebsnummer gekennzeichnet). Besonders kompakt wirken die rote Gruppe der Betriebe 9, 11, 12, 13, 15 und 16 und die lila Gruppe der Betriebe 3, 5, 6 und 17. Die drei gelb gekennzeichneten Betriebe lassen sich im Andrews-Plot keiner der anderen Gruppen gut zuordnen. Übersicht A20 gibt einige Informationen zu dieser Gruppierung, die insgesamt die Wahl der Gruppen bestätigen. Nur zwischen den Betrieben in Gruppe zwei und drei (rot und lila) besteht eine etwas höhere, mittlere Ähnlichkeit zu den Betrieben der anderen als der eigenen Gruppe. Der Parallelkoordinatenplot (A44) läßt die im Andrews-Plot vorgenommene Gruppierung als sinnvoll erscheinen, obwohl deutlich wird, daß die Gruppierung vor allem in den ersten beiden Dimensionen vorliegt und in dritter und vierter Dimension bei weitem nicht so ausgeprägt ist. Diese Beeinflussung des Andrews-Plots durch die ersten Variablen ist bekannt (ROVAN, 1994). Der Andrews-Plot ist also nicht in der Lage eine gleichberechtigte Abbildung aller ausgewählten Variablen (in diesem Fall der Dimensionen) zu erzeugen. Das Trellis-Display (A45) schließlich kann auch nur ein unbefriedigendes Hilfsmittel beim Versuch der Darstellung von vier Dimensionen sein. Tatsächlich wird eine deutliche Entzerrung im Vergleich zu den

zweidimensionalen Plots erreicht, eine intuitive Erfassung der Distanzen zwischen den Betrieben geht aber durch die Konditionierung verloren.

- Eine inhaltliche Interpretation der vier Gruppen wird durch den Parallelkoordinatenplot einiger ausgewählter Variablen in Abbildung A46 unterstützt.

⇒ Die ausgewählten Variablen, die mit Ausnahme des geschätzten Energieverbrauchs (energie) alle zum Bereich 'Platzbedarf und Rücken' gehören, zeigen von allen Variablen die deutlichste Beziehung der gewählten Gruppierung zu den Originaldaten. Bei den anderen Merkmalen gibt es in den meisten Fällen eine noch stärkere Überschneidung zwischen den Gruppierungen. Allerdings lassen auch Andrews- und Pallelkoordinatenplot (A43 und A44) eine sehr klare Gruppierung nicht erwarten. Für den Fall der ausgewählten Merkmale ist aber doch eine recht gute Abgrenzung zum Beispiel der blauen Gruppe möglich. Alle Betriebe dieser Gruppe liegen im Bereich mittleren Energieverbrauchs, kultivieren nur kurze Zeit im, eher engen, Endstand (endstand und woaufend) und haben ausnahmslos nur einmal gerückt (anz_ruec). Ihr Platzzeitwert (net_woqm) liegt daher eher im mittleren und oberen Bereich. Der scheinbare Widerspruch zwischen kurzer Zeit im Endstand und dennoch relativ hohen Platzzeitwertwerten erklärt sich durch das geringe Verhältnis von Aufstellen zu Endstand (auf_end), das heißt die Kultur wurde bereits zu Kulturbeginn weit gestellt und dann nur einmal (relativ wenig) gerückt. Die lila Gruppe hebt sich deutlich von dieser Gruppe ab. Enger gestellt zu Kulturbeginn, weiter im Endstand (das heißt weniger Pflanzen je m^2) ausnahmslos höherer Energieverbrauch, Platzzeitwert und Anzahl Wochen auf Endstand, sowie in allen Fällen mehrfaches Rücken. Der sehr ähnliche Verlauf von grüner und roter Gruppe ist ebenso sichtbar wie der extreme Verlauf von zumindest zwei der gelb gekennzeichneten Betriebe, die sehr eng und somit mit geringen Platzzeitwert und Energieverbrauch kultiviert haben.

3.1.2.4 Variablenset 4 - Ermittlung der Strukturdaten

Bei den Strukturdaten handelt sich um neun, ausnahmslos nominalskalierte, binäre Variablen, wobei die Bildung der beiden Klassen durch den Verfasser nach sachlogischen Gesichtspunkten erfolgt ist (siehe Übersicht A4). Andere Klassenbildungen sind natürlich denkbar und zu rechtfertigen. Die binäre Struktur erlaubt eine gute Übernahme in die Auswertung der Daten als multiple Korrespondenzanalyse der Indikatormatrix.

a-Multiple Korrespondenanalyse und Interpolationsbiplot

- Im ersten Schritt erfolgt die multiple Korrespondenanalyse der Indikatormatrix. Abbildung A47 zeigt die Konfiguration der Betriebe und der Variablen in Normalkoordinaten in getrennten Plots, Abbildung A48 den gemeinsamen Plot in Normalkoordinaten und Abbildung A49 den gemeinsamen Plot in Standard- (Variablen) und Normalkoordinaten (Betriebe). Die Eigenwerte der multiplen Korrespondenzanalyse, die Koordinaten der Merkmale (Standard- und Normalkoordinaten) und der Betriebe (Normalkoordinaten), sowie die Überprüfung der

Interpolation sind in Übersicht A21 widergegeben.

⇒ Abbildung A47a ist zu entnehmen, daß bei der Variablenkonfiguration zwei nahezu orthogonal zueinander liegende Variablengruppen existieren, und zwar auf der einen Seite die Merkmale Bewässerungsverfahren 1 (bw1_f, bw1_k), Stellfläche (sf1_a, sf2_a, sf1_m, sf2_m), Substrate (subs, ee) und Region (ost, west), auf der anderen Seite die Variablen Absatzwege (vm1, vmg1), Betriebsgröße (fw10, fg10), Produktionsmenge (mw50, mg50) und Bewässerungsverfahren 2 (bw2_f, bw2_k). Demnach dominieren bei den beteiligten Betrieben aus dem westlichen Münsterland (wes) Bewässerungsverfahren von unten (bw1_f) auf modernen Stellflächen (sf1_m, sf2_m) und die Verwendung von Einheitserden (ee), während bei den beteiligten Betrieben aus dem östlichen Münsterland (ost) Bewässerungsverfahren von oben (bw1_k) auf traditionellen Stellflächen (sf1_a, sf2_a) und die Inanspruchnahme anderer Substratlieferanten als die der Einheitserden überwiegen (subs). Die kleineren Betriebe (fw10) produzieren größere Mengen (mg50), die über mehr als nur einen Absatzweg vermarktet werden (vmg1), während die größeren Betriebe (fg10) weniger produzieren (mw50) und auf einen Absatzweg (vm1) spezialisiert sind. Der scheinbare Widerspruch zwischen größerer Fläche und geringerer Produktionsmenge mag so interpretiert werden, daß die Cyclamenkultur nach wie vor relativ arbeitsintensiv und schwierig zu mechanisieren ist und daher besser in den Arbeits- und Produktionsablauf des kleineren Produktionsbetriebes als den des Massenproduzenten paßt. Bei den Betrieben in Abbildung A47b deutet sich ebenfalls eine Gruppierung der Betriebe in zwei Gruppen an, und zwar auf der einen Seite die Betriebe 7, 11, 13, 14, 16, 19, und auf der anderen Seite die Betriebe 1, 2, 4, 5, 9, 12, 20. Schließlich fällt erneut Betrieb 3 durch eine besondere, von den anderen Objekten weit entfernte Lage, auf (siehe auch c-). Der gemeinsame Plot in A48 bietet nur eine unbefriedigende Möglichkeit der Interpretation der Beziehungen zwischen Objekten und Merkmalen, da die Distanz zwischen Reihen- und Spaltenkoordinaten nicht definiert ist. Hier bieten die Prediktionsregionen (siehe b-) weit mehr Möglichkeiten. Durch die Darstellung der Betriebe in Normal- und der Variablen in Standardkoordinaten in Abbildung A49 ergibt sich allerdings die Möglichkeit der exakten graphischen Interpolation, das heißt durch diese Darstellung erhält der Korrespondenanalyseplot eine echte Biplot-Interpretation, die der traditionelle Korrespondenanalyseplot (Abbildung A48) nicht besitzt. Im Beispiel bezeichnet die grüne Linie, die die Merkmale von Betrieb 3 miteinander verbindet, die Interpolationsregion für diesen Betrieb, der am Zentroid des entstandenen Polygons liegt. Übersicht A21 zeigt, daß die Interpolation exakt ist. Auf diese Art und Weise lassen sich auch Objekte in den Plot interpolieren, deren Merkmale bekannt sind, die aber nicht an der Konstruktion des Plots beteiligt sind (zum Beispiel nachträglich aufgenommene Objekte).

b-Multiple Korrespondenanalyse und Prediktionsbiplot

– Aus den CLPs (category level points), das heißt den Koordinaten der Variablen, lassen sich in der Korrepondenzanalyse Prediktionsregionen bilden, die das diskrete Gegenstück zu den

Biplot-achsen oder Biplotbahnen der linearen und nichtlinearen Biplots darstellen. Abbildung A50 beinhaltet diese Regionen für einzelne Variablengruppen bei Verwendung des üblichen Distanzmaßes in der Korrespondenzanalyse, der Chi-Quadrat-Distanz (mca), Abbildung A52 für einzelne Variablen bei Verwendung des Extended Matching-Koeffizienten (emc, eine Variante des Simple Matching-Koeffizienten nach GOWER & HAND, 1996). In den Abbildungen A51 und A53 erfolgt die gemeinsame Darstellung aller Variablen bei Verwendung der beiden Distanzmaße. Eine Prediktionsregion wird durch eine mit der entsprechenden Farbe gezeichneten Linien abgegrenzt und durch ein Kürzel oder einen Text beschrieben. Die Fehlprediktionen, das heißt die Lage von Objekten in einer durch die CLPs bestimmten Region zu der sie in Wirklichkeit nicht gehören, die natürlich auch in dieser Art der Darstellung nicht vermieden werden können, sind in Übersicht A22 zusammengefaßt.

⇒ Beide Distanzmaße produzieren sehr ähnliche Repräsentationen der Daten, sowohl im Bezug auf die Lage der Betriebe als auch hinsichtlich der Definition der Prediktionsregionen. Eine besonders gute Prediktion wird für die Variablen Produktionsmenge und Stellfläche erreicht, besonders unscharf ist die Klassenvorhersage bei den Merkmalen Substrate und Absatzwege. Insgesamt führt die mca-Prediktion zu einer Fehlprediktion durch die Prediktionsregionen in 23 (von 140 Fällen), die emc-Prediktion zu einer Fehlprediktion von 19 (von 140 Fällen). Besonders häufig fehleingeordnet werden die Betriebe 1 und 6 (bei jeweils drei Merkmalen bei mca und zwei Merkmalen bei emc). Trotz dieser Einschränkungen liefern die Prediktionsbiplots sehr kompakte Zusammenfassungen der in den Daten enthaltenen Informationen. So grenzt sich in A51 deutlich eine Gruppe der Betriebe, die auf herkömmlichen Stellflächen mit über-Kopf Bewässerung und anderen Substraten als der Einheitserde kultivieren (Betriebe 3, 7, 8, 11, 13, 14, 16, 18), von einer zweiten Gruppe (Betriebe 1, 2, 4, 5, 10, 12, 15, 17, 20) mit den entgegengesetzten Merkmalen ab. Da darüberhinaus in A51 auch die Regionen für Produktionsmenge und Betriebsgröße widergegeben sind, lassen sich diese Gruppen noch detaillierter beschreiben, so zum Beispiel die Gruppe der Betriebe 7, 11, 13, 14, 16, und 19, die zur ersten Gruppe hinsichtlich der Merkmale Stellfläche, Bewässerungssystem und Substrate zählen und darüberhinaus relativ große Mengen (über 50000 Stück), bei relativ kleinerer Betriebsgröße (unter 10000 m^2) kultuivieren. Allerdings ist auch hier zu beachten, daß es in einzelnen Fällen zu Fehlrepräsentationen durch die Dimensionserniedrigung kommt.

⇒ Um derartige Fehlrepräsentationen zu vermeiden, bietet sich die alternative Darstellung mit Hilfe beschrifteter Objektmeßwerte-Plots, wie sie in Abbildung A54 zu sehen sind, an. Natürlich sind dann, an Stelle von einer Abbildung, in diesem Fall bei Beschriftung durch alle Variablen mindestens sieben oder besser acht Abbildungen erforderlich. A54 beinhaltet nur fünf der beobachteten Merkmale, sowie in einer Teilabbildung die Beschriftung der Punkte mit der Betriebsnummer.

c-Prüfung der internen Stabilität

– Es soll nun die Frage untersucht werden, wie stabil die Repräsentation der Betriebe und ihrer Strukturmerkmale im Korrespondenzanalyseplot ist.

* Dazu wird folgendes Verfahren, das sich an die Ausführungen von GREENACRE, 1984, anlehnt gewählt. Im ersten Schritt erfolgt die Analyse der Datenmatrix jeweils sukzessive ohne Betrieb 1, 2, 3 und so weiter. Für jedes der reduzierten Datensets werden Reihen- und Spaltenkoordinaten mit Hilfe der Korrespondenzanalyse berechnet. Um der willkürlichen Vergabe der Vorzeichen zu begegnen, wird im zweiten Schritt eine Prokrustes-Rotation mit der Konfiguration der vollen Datenmatrix als fixer Konfiguration durchgeführt (ohne Dilation, jedoch bei erfolgter Standardisierung (auf Sum of Squares = 1) und Zentrierung (am Ursprung)), die zu einer größtmöglichen Deckung der Konfiguration der vollen Datenmatrix mit den reduzierten Datenmatrizen führt. Während dies bei der Berechnung der Variablen keine Probleme (Verwendung der Konfiguration aller Objekte) bereitet, führt die Elimination eines Objektes natürlich dazu, daß auch nur eine reduzierte Datenmatrix als fixe Ausgangsmatrix im Fall der Berechnung der Objektkonfiguration angenommen werden kann, das heißt, aus den Ergebnissen der Korrespondenzanalyse der vollen Datenmatrix, muß jeweils das entsprechende Objekt entfernt werden, bevor die Rotation erfolgt. Der besseren Übersichtlichkeit halber werden dann im dritten Schritt Plots erstellt, die die äußersten Punkte aller Konfigurationen mit konvexen Hüllen verbinden, so daß ein Eindruck davon entsteht, in welchem Ausmaß sich die einzelnen Punkte im zweidimensionalen Koordinatensystem bewegen. Während bei der Darstellung der Hüllen der Variablen die Orginalkoordinaten den Variablenpunkt im Plot bezeichnen, wird bei der Darstellung der Betriebe der mittlere Koordinatenwert aus den rotierten Konfigurationen gewählt.

⇒ Abbildung A 55 zeigt zunächst einige Dotplots der Residuen bei der Prokrustes-Rotation bei Elimination der Betriebe 1, 2 und 3. Die Konfigurationen scheinen in diesen Fällen sehr stabil zu sein. Allerdings ergibt sich bei Entfernen von Betrieb 3, vor allem bei Betrachtung der Variablen, eine auffällige Veränderung. Diese ist bedingt dadurch, daß nur für Betrieb 3 das Merkmal bw2_k zutrifft, das heißt Betrieb 3 der einzige Betrieb ist, der auf beiden Stellflächen von oben bewässert (bw2_k). Die Abbildungen A56 und A57 verdeutlichen jedoch, daß insgesamt die interne Stabilität der Repräsentationen der Merkmale und der Objekte recht hoch ist, das heißt keiner der Betriebe übt einen die Abbildungen wesentlich verändernden Einfluß aus. Allerdings fällt bei der Betrachtung der Variablen in Abbildung A56 schon auf, daß einzelne Merkmale sich überschneidende Felder besetzen, insbesondere die Merkmale Produktionsmenge und Betriebsgröße (mg50, fw10, mw50, fg10). Außerdem ist die sehr große konvexe Hülle beim Merkmal bw2_k zu erkennnen, die auf der Elimination von Betrieb 3 beruht. Wird Betrieb 3 entfernt, verschiebt sich dieser Merkmalspunkt an den Ursprung. A57 zeigt, daß es bei den Objektkonfigurationen nur sehr geringe Unterschiede durch Entfernen eines einzelnen Betriebes gibt, die interne Stabilität der Objektkoordinaten also sehr hoch ist.

3.1.2.5 Vergleich aller Variablensets

Nachdem in den vorangegangenen Abschnitten die einzelnen Variablensets separat analysiert worden sind, wird nun eine gemeinsame Untersuchung der unterschiedlichen Datensätze durchgeführt. Zunächst erfolgt eine objektbezogene Betrachtung durch eine multiple Procrustes-Analyse und dann eine variablenbezogene Analyse mit Hilfe der generalisierten kanonischen Analyse.

a-Multiple Prokrustes-Analyse

* Die multiple Procrustes-Analyse erzeugt eine Matrix der Quadratsummen der Abweichungen (Residuen) beim paarweisen Vergleich aller Konfigurationen, wobei eine Dilation zugelassen und eine Normierung der auf gleiche Varianzen der Ausgangskonfigurationen vorgenommen wird. Als Ausgangskonfigurationen werden für die Beurteilung der Qualitätsmerkmale die Objektkoordinaten der Korrespondenzanalyse der bipolaren Daten verwendet (3.1.2.1); bei Substratanalysewerten und Kulturmaßnahmen (3.1.2.2 und 3.1.2.3) werden die Objektkoeffizienten nach einer Hauptkoordinatenanalyse eingesetzt (Proximitätsmatsmaß allgemeiner Ähnlichkeitskoeffizient); die Strukturmerkmale schließlich gehen in Form der Objektkoordinaten der Analyse der Indikatormatrix ein (3.1.2.4, Basis Chi-Quadrat-Distanz). Die symmetrische Matrix der Residuen wird dann mit einer Hauptkoordinatenanalyse und einer ordinalen, mehrdimensionalen Skalierung (50 Startkonfigurationen, primäre Behandlung gebundener Werte, least squares stress) verrechnet.

– Abbildung A 58 zeigt zwei Dshade-Diagramme der Matrix der quadrierten Residuen, ausgedrückt als Ähnlichkeitswerte, A58a in aufsteigender Reihenfolge der laufenden Konfigurationsnummer, A58b etwas umsortiert, um einzelne Ähnlichkeitsgruppierungen stärker hervorzuheben. Die Abbildungen A59 und A60 beinhalten die Konfigurationen der Hauptkoordinatenanalyse und der ordinalen mehrdimensionalen Skalierung in den ersten beiden Dimensionen, mit und ohne überlagerten Multiple Spanning Tree. Übersicht A23 schließlich liefert Koordinaten- und Eigenwerte, sowie weitere ergänzende Informationen zur Analyse.

⇒ Die Dshade-Diagramme, die farbig codiert die Matrix der Residuen wiedergeben, verdeutlichen, daß die Ähnlichkeiten zwischen den einzelnen Konfigurationen insgesamt gesehen relativ gering sind (maximale Ähnlichkeit 0,336). Es fällt auf, daß die Matrix der Strukturmerkmale den übrigen Konfigurationen am nächsten liegt, während die Konfiguration der Schattiersollwerte zu allen übrigen Konfigurationen nur sehr geringe Ähnlichkeiten besitzt. Diese Beobachtung wird auch durch die zweidimensionalen Plots in A59 und A60, die eine recht gute Repräsentation der Distanzen liefern (keine auffälligen Überschneidungen im Multiple Spanning Tree) verstärkt. Die zentrale Stellung der Konfiguration der Strukturmerkmale wird ebenso sichtbar (siehe auch geringste Zentroid Distanz in Übersicht A23), vor allem im Plot der ordinalen, mehrdimensionalen Skalierung. Sowohl der Plot der Hauptkoordinatenanalyse als auch der Plot

der ordinalen mehrdimensionalen Skalierung lassen den Schluß zu, daß die ‘Sierra’ Konfigurationen der Qualitätsbonituren den übrigen Konfigurationen (Substratanalysewerte, Kulturmaßnahmen, Strukturmerkmale) ähnlicher sind als die ‘Concerto’-Konfigurationen der Qualitätsbonituren. Eine eindeutige Gruppierung ist jedoch nicht erkennbar. Deren Fehlen, die insgesamt gesehen niedrigen Ähnlichkeitswerte und die relativ gleichartigen Zentroid-Distanzen führen vielmehr zu der Schlußfolgerung, daß alle Variablensets doch recht unterschiedliche Konfigurationen erzeugen, daß also die relative Lage der Objekte in den einzelnen Variablensets recht unterschiedlich ist und es sich nur schwerlich aus dieser Analyse ableiten läßt, daß zum Beispiel Betriebe, die gleichartig kultiviert haben auch gleichartige Qualitätsergebnisse erbracht haben. Erfolg oder Mißerfolg in der Kultur kann also demnach kaum durch die bestimmten Kulturmerkmale in ihrer Gesamtheit hergeleitet werden, was natürlich nicht besagt, daß eine solche Beziehung nicht doch in einzelnen Punkten besteht (bisweilen ist ja eine solche Beziehung zu vermuten, siehe zum Beispiel 3.1.2.3 a-).

b-Generalisierte kanonische Analyse (OVERALS)

* In der generalisierten kanonischen Analyse geht es um den detaillierten Vergleich der unterschiedlichen Variablensets und zwar in der Form, daß jeweils sechs Variablensets der Analyse übergeben werden. Dabei werden fünf Datensets immer verwendet (Strukturdaten, Substratanalysewerte, Platzbedarfsmerkmale, Temperaturführung und Hauptkomponentenwerte der Schattiersollwerte) und durch ein weiteres Qualitätsmerkmale-Variablenset ergänzt, das heißt sukzessive durch die Qualitätsbeurteilungen von ‘Sierra’ in Woche 44 und ‘Sierra’ in Woche 48 sowie von ‘Concerto’ in Woche 44 und ‘Concerto’ in Woche 48. Die Aufteilung in diese Variablensets beruht auf einer subjektiven Entscheidung und es durchaus zulässig, andere Zuordnungen zu wählen, die dann dementsprechend auch Auswirkungen auf die Analyseergebnisse haben können.

– Die gewählten Meßniveaus der einzelnen Variablen, und ihre Transformationen, wo erforderlich[38], sind in Übersicht A24 zusammengefaßt. Übersicht A25 beinhaltet die Loss-Werte der vier Analysen, die Übersichten A 26 und A27 die multiplen Anpassungswerte in zwei Dimensionen in den verwendeten Variablensets. In den Abbildungen A 61 bis A64 sind die Komponentenladungen der Variablen, der Übersichtlichkeit halber getrennt, wiedergegeben. A65 schließlich zeigt beispielhaft den Komponentenladungenplot aller Variablen für die Variante mit ‘Concerto’ Woche 44.

⇒ Übersicht A25 zeigt, daß in allen vier Analysen eine sehr gute Anpassung erreicht wird und die

[38] Die in dem Zusammenhang mit der generalsierten kanonischen Analyse verwendete SPSS-Prozedur OVERALS akzeptiert nur Variablenwerte kleiner 100; liegen die Variablenwerte darüber, werden sie automatisch in Ränge transformiert.

Loss-Werte insgesamt gesehen gering sind. Nur das Strukturdatenset hat in allen Analysen vergleichsweise hohe Loss-Werte. Darüber hinaus wird durch die Eigenwerte der ersten beiden Dimensionen deutlich, daß die erste und die zweite Dimension in allen Fällen nahezu in gleichem Umfang an der Repräsentation der Variablenbeziehungen beteiligt sind. Den Übersichten A26 und A27 ist die Diskriminationsstärke der einzelnen Variablen zu entnehmen. Multiple Anpassungswerte von größer 1 sind farbig hervorgehoben. Auffällig ist die Unterschiedlichkeit dieser stark diskriminatorischen Variablen in den vier Analysen, so daß sich demnach keine durchgängige Interpretation der Beziehungen der Variablensets untereinander ableiten läßt. Auch wird in keinem Fall eine oder beide Dimensionen durch ein Variablenset in besonderem Ausmaß beeinflußt. Allerdings gehören die Variablen Endstand und die erste Hauptkomponente der Schattiersollwerte (das heißt der Maßstab für die mittlere Schattierung während der gesamten Kultur) in allen vier Analysen zu den Merkmalen, die besonders große Unterschiede zwischen den Betrieben aufweisen. Demgegenüber weisen zum Beispiel die Substratanalysewerte in Woche 41 weder bei ‘Sierra’ noch bei ‘Concerto’ eine multiple Anpassung von über 1 aus. Es gibt also eine Vielzahl von Variablen, die nur sehr schlecht repräsentiert werden.

⇒ An Hand der Plots der Komponentenladungen in den Abbildungen A63 bis A66 soll nun der Versuch unternommen werden, die Beziehungen zwischen den Variablensets, insbesondere im Hinblick auf die Qualitätsbeurteilungen, zu interpretieren. Auffällig an allen Plots ist, daß die einzelnen Variablen der verschiedenen Datensets fast immer alle Räume der Abbildungen besetzen, daß sich also die Variablensets nicht eindeutig gruppieren. Eine recht gut deutbare Abbildung liefern jedoch die Plots für ‘Concerto’ in Woche 44 (einzeln in Abbildung A63, überlagert in Abbildung A65). Die Qualitätsmerkmale Gesamteindruck, Knospenbesatz, Vergilbung und Welke besetzen einen Bereich, der auch stark von Platzbedarfsvariablen belegt ist, und zwar von Wochen auf Endstand, Verhältnis Aufstellen zu Endstand, Anzahl Rückvorgänge und Platzzeitwert, und in dem sich die Stellflächenklassifizierung (moderne Stellflächen) und Substratwahl (Einheitserden) wiederfinden. Mit guten Qualitäten korrelieren somit vor allem die Pflanzen von Betrieben, die lange auf Endstand kultiviert werden, deren Verhältnis von Aufstellen zu Endstand hoch ist (das heißt, die zuerst sehr dicht und dann recht weit kultiviert werden), die häufig gerückt und die insgesamt mit einem hohen Platzbedarf kultiviert und zudem auf modernen Stellflächen produziert werden. In diesem Beispiel finden sich demnach bekannte Beratungs-empfehlungen zur optimalen Kulturführung bei Cyclamen wieder, die rechtzeitiges Rücken und das Schaffen der jeweils optimalen Standweite betonen. Demgegenüber steht ein Variablencluster, in erster Linie gebildet aus Temperaturdaten und Substratanalysewerten, so daß demnach hohe Temperaturen (vor allem Lüftungstemperaturen) und hohe Nährstoffgehalte eher mit minderen Qualitäten in Zusammenhang stehen. In diese Variablengruppe fallen auch die Hauptkomponentenwerte der Schattiersollwerte (erste Hauptkomponente) und bestätigen damit, daß auch eine starke Schattierung mit geringen

Qualitäten korrespondiert. Auch an dieser Stelle bestätigen sich Kulturhinweise bei Cyclamen, die eine helle und luftige Kulturführung bei mittlerer Düngung empfehlen (siehe zum Beispiel zur Cyclamenkultur bei HASS-TSCHIRSCHKE, 1994 oder HORN, 1996). Nahezu orthogonal zu diesen Bereichen befinden sich an entgegengesetzten Polen und fast vollständig in der zweiten Dimension auf der einen Seite die Beurteilung des Knospenbesatzes und auf der anderen Seite die Variablen end (Endstand) und licht2 (zweite Hauptkomponente der Schattiersollwerte). Dies führt zu der Interpretation, daß sehr weiter Endstand und eine im Kulturverlauf gleichbleibende oder zunehmende Schattierung, bei ‘Concerto’ vor allem ungünstige Auswirkungen auf den Knospenbesatz hat.

⇒ Die einzelnen Abbildungen in A66 illustrieren, daß die im vorangegangenen Abschnitt (Abbildung A63 und A65) angesprochenen, mit der generalisierten kanonischen Analyse herausgearbeiteten Beziehungen zwischen Kulturbedingungen und Produktqualität tatsächlich auf einige der wesentlichen Variablenbeziehungen in der Betrachtung mit ‘Concerto’ Woche 44 hinweisen, und daß sie sich durch die Betrachtung der Orginalwerte nachvollziehen läßt. Beim Qualitätsmerkmal Vergilbung zum Beispiel, wird sichtbar, daß alle Betriebe mit modernen Stellflächen eine Boniturnote 7 oder besser besitzen und alle Betriebe auf herkömmlichen Stellflächen, mit einer Ausnahme, eine Boniturnote von 7 oder schlechter aufweisen (A66a). Ähnliches gilt für die Substratwahl (A66b). Beim Qualitätsmerkmal Krankheitsbefall fällt auf, daß sehr gute Bonituren (Noten 8 und 9) nur von Betrieben mit K-Werten von unter 200 mg/l Substrat (bei Messung in der Kulturmitte, Woche 29) erzielt werden (A66c). Weniger gut ist die Beziehung von Krankheitsbefall und Schattierung in A66d nachvollziehbar.

⇒ Entsprechende Interpretationsansätze können auch auf die übrigen Plots der Komponentenladungen angewendet werden. Allerdings sind die Beziehungen der Variablensets in den Abbildungen A61, A62 und A64, weniger deutlich als in A63 und es ergibt sich kein einheitliches Bild für die vier Analysen, was erneut die Unterschiedlichkeit der Sorte-Woche-Kombinationen, auch im Bezug auf ihre Beziehung zu den übrigen bestimmten Merkmalen, unterstreicht. Aufgrund der vielfach geringen Korrelationen zwischen den Merkmalen innerhalb und zwischen den einzelnen Variablensets kann in dieser betriebsbegleitenden Untersuchung mit Cyclamen nicht auf eindeutige Wirkungszusammenhänge zwischen Kulturbedingungen und Produktqualität geschlossen werden.

3.2 Kennzahlen des Kennzahlenvergleichs für Topfpflanzenbetriebe des Bundesgebietes der Jahre 1992 bis 1994

3.2.1 Einführung

Seit 1957 veröffentlicht der Arbeitskreis Betriebswirtschaft im Gartenbau e.V., Hannover, Kennzahlen für den Betriebsvergleich. Es handelt sich bei den Kennzahlen um durch den Arbeitskreis aufgearbeitete Daten aus den steuerlichen Jahresabschlüssen gartenbaulicher Betriebe (AKBWL, 1996). Das entwickelte Kennzahlensystem wird von der Beratung in erster Linie in der einzelbetrieblichen Beratung eingesetzt, das heißt der Berater versucht die wirtschaftliche Situation des Betriebes mit Hilfe der Kennzahlen zu analysieren, um Schwachstellen oder Stärken aufzudecken. Vertikaler und horizontaler Betriebsvergleich führen zu einer Einschätzung der Entwicklung der Betriebsergebnisse über mehrere Jahre (vertikal) und zu einer Abgleichung der Ergebnisse des betroffenen Betriebes mit den Ergebnissen anderer, gleichgelagerter Betriebe (horizontal) (STORCK & BOKELMANN, 1995). Die Zweckmäßigkeit und Vergleichbarkeit vieler Kennzahlen ist nicht unumstritten und bisweilen werden durch die Beratung aus den Kennzahlen des Arbeitskreises weitere Kennwerte berechnet, die eine bessere Interpretierbarkeit liefern sollen (KÜHNE, 1997). Darüber hinaus ist in der einzelbetrieblichen Beratung auch die Kenntnis betrieblicher Gegebenheiten sehr wichtig, da es sonst, bei ausschließlicher Betrachtung der Zahlen, leicht zu Fehldeutungen kommen kann (STORCK & BOKELMANN, 1995). Die Frage, wie sinnvoll einzelne Kennzahlen oder gar der Kennzahlenvergleich überhaupt ist, wird in dieser Arbeit nicht diskutiert. Einen Überblick über Geschichte und Stand der Forschung im Bereich der Bilanz- und Kennzahlenanalyse im Allgemeinen und im Gartenbau im Speziellen gibt BITSCH, 1994.

Neben der Verwendung als einzelbetriebliches Beratungsinstrument werden die Kennzahlen des Arbeitskreises Betriebswirtschaft gerne verwendet, um allgemeine Wirkungszusammenhänge zu untersuchen, obwohl es sich bei den am Betriebsvergleich beteiligten Betrieben nicht um eine repräsentative Stichprobe des deutschen Gartenbaus handelt. Der große Umfang der erhobenen und ausgewerteten Betriebe (ungefähr 2500 Betriebe pro Jahr aus allen Sparten des Gartenbaus) macht die Daten des Arbeitskreises Betriebswirtschaft aber dennoch interessant, da es sich um die umfassendste und aktuellste veröffentlichte Information zur wirtschaftlichen Situation von Gartenbaubetrieben handelt. BOKELMANN, 1993, sucht in den Kennzahlen nach Informationen über den Erfolg und die finanzielle Situation der Betriebe, um Schlüsse auf Gefährdungen und Chancen der Betriebe zu ziehen. Er bedient sich dabei verschiedener Verfahren der Diskriminanzanalyse, um Prognosemodelle aus den Kennzahlen abzuleiten. BITSCH, 1994, nutzt im wesentlichen die Clusteranalyse, um homogene Betriebsgruppen zu finden, die dann durch einzelne Kennzahlen beschrieben werden und zu einer Trennung in erfolgreiche und weniger erfolgreiche Betriebe führt. Ähnlich geht GOTTSCHLICH, 1995, im landwirtschaftlichen Bereich vor. Er bedient sich ebenfalls überwiegend der Clusteranalyse, um Schwachstellenprofile zu identifizieren und besonders aussagekräftige Kennzahlen für die Schwachstellenanalyse herauszufiltern. Ziel der vorliegenden Arbeit ist nicht die Diskussion der bereits erarbeiten

Ergebnisse oder die Bildung weiterer Gruppen, Klassifizierungen oder Modelle, sondern die Diskussion graphischer Analysemöglichkeiten, die die Zusammenhänge zwischen den betriebswirtschaftlichen Kennzahlen auf vielfältige Weise transparent machen.

Es wird in fünf Schritten vorgegangen. Zunächst erfolgt eine erstes ‘Anschauen’ der Daten, um ein Gefühl für Lage- und Dispersionsparameter einiger Kennzahlen zu gewinnen, und um mögliche Korrelationen aufzudecken (3.2.2.1). Im zweiten Abschnitt wird dann überprüft, ob den traditionell im Kennzahlenheft verwendeten Klassifizierungen (Betriebsgröße, regionale Lage, Jahre) unterschiedliche Quellen der Variabilität zugrunde liegen (3.2.2.2). Die Auseinandersetzung mit verschiedenen Gruppierungs- und Segmentierungsverfahren erfolgt im dritten Abschnitt (3.2.2.3) Anschließend werden diskrete graphische Modelle als Weg der Ermittlung von Zusammenhängen zwischen Kennzahlen vorgeschlagen (3.2.2.4). Den Abschluß bildet eine Umsetzung des Kennzahlenkatalogs in interaktive Liniendiagramme der formalen Begriffsanalyse, die die intuitive Untersuchung der ermittelten Daten unterstützt (3.2.2.5). Die vorgelegten Analysen berücksichtigen also einerseits Anwendungen, die für die einzelbetriebliche Beratung relevant sind und führen andererseits zu einer neuen Betrachtungsweise allgemeiner Fragestellungen.

Bei den verwendeten Daten handelt es sich um ausgewählte Kennzahlen der Topfpflanzenbetriebe des gesamten Bundesgebietes aus den Jahren 1992, 1993 und 1994 (UHTE, 1997). Auf die Herleitung und Bedeutung der einzelnen Kennzahlen wird hier nicht eingegangen. Sie sind zum Beispiel in AKBWL, 1996, ausführlich beschrieben.

3.2.2 Darstellung der Ergebnisse

Es erfolgt nun die Analyse ausgewählter Kennzahlen der Jahre 1992 bis 1994 in den beschriebenen fünf Schritten. Die Vorgehensweise entspricht der von Kapitel 3.2.1, das heißt, daß zunächst eine oder mehrere Abbildungen oder Analyseschritte kurz erläutert (gekennzeichnet durch ein vor den Absatz gestelltes „ - "-Zeichen) und dann interpretiert (gekennzeichnet durch ein vor den Absatz gestelltes „⇒"-Zeichen) werden. Spezielle methodische Anmerkungen folgen auf ein „∗"-Zeichen. Die Abbildungen und Übersichten sind im Anhang Teil I B und II B hinterlegt.

3.2.2.1 Einführende Datenanalyse

- Der Kennzahlenkatalog beinhaltet insgesamt 87 Kennzahlen, die vom Arbeitskreis Betriebswirtschaft in 14 Kategorien eingeteilt werden (Aufwandsstruktur, Unternehmenserfolg, Betriebserfolg, Kapitalstruktur und andere). Aus den 87 Kennzahlen werden für die vorliegende Arbeit 26 Kennzahlen und 9 Gruppierungsvariablen ausgewählt, die in Übersicht B1 zusammengefaßt sind, und die je nach Fragestellung verwendet werden. Bei derAuswahl der Kennzahlen wird im wesentlichen dem Vorschlag von BITSCH, 1994 (Seite 132 f.)[39] gefolgt. Die Übersichten B2 bis B6 informieren über univariate Statistiken dieser Variablen. Übersicht B7 beinhaltet Informationen zu den Shingles (nach dem equal count-Algorithmus, siehe 2.5.1.3) für die Kennzahlen Glasfläche, Anzahl Arbeitskräfte und Unternehmensertrag. Betrachtet werden ausschließlich die Kennzahlen derjenigen Topfpflanzenbetriebe, die in jedem der drei vorliegenden Jahre am Betriebsvergleich teilgenommen haben, das sind 297 Betriebe pro Jahr und 891 Fälle insgesamt. Bei allen Kennzahlen handelt es sich um verhältnis- oder intervallskalierte Variablen.

⇒ Beim Vergleich der Klassifizierungen der Betriebe nach Anzahl der Arbeitskräfte, Glasfläche und Unternehmensertrag nach dem Schema des Arbeitskreises Betriebswirtschaft (Übersicht B2) und der gebildeten Shingles (Überlappungsbereich 10%) (Übersicht B7) wird eine erheblich voneinander abweichende Bildung der Klassengrenzen deutlich und zwar nach oben gerichtet, das heißt, um gleich große Klassen zu bilden ist eine Verschiebung der Klassengrenzen nach oben erforderlich. Eine noch etwas stärkere Differenzierung bietet die Bildung von vier Klassen. Diese vier Klassen, gebildet mit Hilfe des equal-count-Algorithmus und einem Überlappungsbereich von 10%, liegen den weiter unten besprochenen Trellis-Displays zugrunde. Die univariaten Statistiken in den Übersichten B3 bis B6 zeigen Auffälligkeiten vor allem durch ihre starke Streuung, sowie die hohen Werte bei Schiefe und Kurtosis und eine in

---

[39] Auf eine eigene methodisch-begründete Variablenselektion wird mit Blick auf die Ziele der vorliegenden Untersuchung verzichtet. Der Ausschluß einiger Kennzahlen und die Hinzunahme anderer, von BITSCH, 1994, nicht ausgewählter Kennzahlen, ist lediglich durch das Interesse oder Desinteresse des Verfassers an der Betrachtung einzelner Kennwerte begründet.

vielen Fällen erhebliche Abweichung zwischen Mittelwert und Median. Dies deutet auf Ausreißer oder extreme Werte, sowie erhebliche Abweichungen von der Normalverteilung hin.

– Diese Abweichungen von der Normalverteilung gilt es näher zu betrachten. Für einige ausgewälte Variablen zeigen die Abbildungen B1 und B2 Boxplot, Histogram und Normal-q-q-Plot in einer Abbildung. Formale Normalverteilungstests sind in Übersicht B8 zusammengefaßt. Wie sich der Ausschluß einiger besonders extremer Werte einiger univariate Statistiken auf einige Kennzahlen auswirkt, ist Übersicht B9 zu entnehmen.

⇒ Die Graphiken lassen in allen Fällen mehr oder weniger stark ausgeprägte Abweichungen von der Normalverteilung erkennen, wobei diese vor allem auf extremen Werten von einigen Betrieben am oberen oder unteren Ende (oder an beiden Enden) der jeweiligen Skala beruhen. Das heißt, das es neben einer Vielzahl ‘mittlerer’ Betriebe, immmer eine gewisse Anzahl von Betrieben gibt, die mit besonders hohen oder besonders niedrigen Werten in den jeweiligen Kennzahlen auftauchen. Der Ausschluß einzelner Fälle führt zu einer gewissen Annäherung an die Normalverteilung, kann aber in vielen Fällen die vorhandene Schiefe nicht beseitigen. Für die Darstellung in den Trellis-Displays (siehe unten) reicht aber schon diese Entfernung der sehr extremen Fälle aus, um in den Darstellungen etwas erkennen zu können (es handelt sich um die in Übersicht B9 aufgeführten missing cases). Eine Vielzahl von Transformationen ist natürlich darüber hinaus denkbar, um die Anpassung an die Normalverteilung zu verbessern. In einzelnen Fällen ist eine Transformation möglicherweise sogar inhaltlich begründbar (zum Beispiel log(Unternehmensertrag)), bei anderen Kennzahlen ist es aber eher fraglich, ob eine Transformation zu einer aussagekräftigeren Variablen führt. Da zudem die meisten der in der Folge eingesetzten Verfahren keine Verteilungsannahmen machen und Transformationen größtenteils inhaltlich nicht begründet werden können, wird auf deren Einsatz, nur zur Erlangung der Normalverteilung, verzichtet. Allerdings erfolgen weiter unten Transformationen in Form der Diskretisierung der kontinuierlichen Variablen. Sie werden an gegebener Stelle diskutiert.

– Die Übersichten B10 und B11 beinhalten die Spearman-Rangkorrelationen für die ausgewälten 26 Kennzahlen, wobei Übersicht B10 die Korrelationen der Variablen untereinander und Übersicht B11 die Korrelationen der Struktur-, Vermögens- und Aufwandsdaten mit den Erfolgsdaten zeigt. Korrelationen von mehr als 0,7 sind durch Fettdruck hervorgehoben.

∗ Diese Hervorhebung wird gewählt, da erst bei einer Korrelation von über 0,7 davon ausgegangen werden kann, daß wenigstens die Hälfte der gesamten Streuung durch die errechnete Korrelation zwischen den beiden betrachteten Variablen erklärt wird (da Bestimmtheitsmaß = r^2). Die meisten Korrelationen sind hier signifikant, ihre absolute Größe ist aber in vielen Fällen gering.

⇒ Sehr hoch miteinander korrelieren Betriebseinkommen/AK, die Reinertragskennzahlen (rtak, rteqm und rtp), Rentabilitätskoeffizient und Reinertragsdifferenz. Aber auch die Korrelationen mit

den anderen Erfolgsvariablen sind vergleichsweise bedeutsam. Demgegenüber sind die Korrelationen innerhalb der anderen Datensets geringer, außer zum Beispiel in so naheliegenden Fällen wie der Korrelation zwische ak und fremdakp, oder der Korrelation zwischen eqm und glasqm. Die Korrelationen zwischen den Erfolgskennzahlen und den Variablen der übrigen Datensets weisen erstaunlicherweise in keinem Fall eine Korrelation von über 0,7 aus, was darauf hindeutet, daß es zumindest in dieser groben Betrachtung nur geringe Hinweise auf klare lineare Beziehungen zwischen zum Beispiel Aufwands- und Erfolgskennzahlen gibt, obwohl schon auffällt, daß fast alle Aufwandskennzahlen zu fast allen Erfolgskennzahlen negative Korrelationen aufweisen[40].

– Dennoch soll ein weiterer Blick in die Beziehungsgefüge durch deskriptive Mittel unternommen werden. Abbildung B3 beinhaltet vier Trellis-Displays mit Boxplots, wobei die konditionierenden Variablen die Betriebsgröße (nach Shinglebildung, siehe Übersicht B7) und die regionale Lage sind. Region 1 ist die Region mit den meisten am Betriebsvergleich beteiligten Betrieben, während die übrigen Regionen die Betriebe aus dem restlichen Bundesgebiet zusammenfassen. Die Abbildungen B4 bis B7 zeigen ebenfalls Trellis-Displays für einige ausgewählte Kennzahlenbeziehungen, die in den Spalten nach den Erfassungsjahren und in den Zeilen nach den Shingles für die Glasfläche beziehungsweise für die Anzahl an Arbeitskräften konditioniert sind. In den einzelnen Panels erscheint die Linie einer Loess-Regression, die den im Punkteschwarm - der selbst nicht abgebildet wird - vorhandenen Trend sichtbar machen soll. Gewählt wird für die Loess-Regression ein Glättungsparameter von 2/3, es findet eine lokal-lineare Anpassung statt, und die Schätzung erfolgt unter iterativer Einbeziehung der Residuen. Die Ziffern hinter ‘n = ‘ geben an, wieviele Einzelwerte hinter dem jeweiligen Boxplot stehen, eine kleine Zusatztabelle, in derselben Sortierung wie die Panels im Trellis-Display, gibt Auskunft über die Anzahl Fälle in den Abbildungen mit den Loess-Regressionslinien. Die blauen Referenzlinien in den Boxplots stehen am Median der betreffenden Kennzahl.

⇒ In allen Boxplots in Abbildung B3 fällt sowohl ein gewisser Effekt durch die Betriebsgröße als auch ein Unterschied zwischen den Regionen auf. Beim Rentabilitätskoeffizienten liegt der Median der Betriebe in Region 1 immer in allen Größenklassen über den Medianen der übrigen Regionen. Zudem nimmt der Median des Rentabilitätskoeffizienten mit der Betriebsgröße zu. Diese Zunahme mit der Betriebsgröße ist auch bei der Kennzahl Lohn je entlohnte AK zu beobachten. Der Effekt ist allerdings in Region 1 wesentlich ausgeprägter als in den übrigen Regionen. Bei der Betrachtung der Kennzahl Heizmaterial je qm Glasfläche wird deutlich, daß in Region 1 weniger an Energiekosten je Quadratmeter im Mittel entstehen, aber kaum ein Effekt der Betriebsgröße zu erkennen ist, während in den übrigen Regionen eine Abnahme der

---

[40] Außer kapkoef, das entspricht aber derselben Beziehungsrichtung da kapkoef = Kapital / Betriebseinkommen.

Heizkosten und deren Streuung mit zunehmender Betriebsgröße auffallen. Bei qm Glasfläche/AK ist es wiederum die Region 1, die einen stärkeren Betriebsgrößeneffekt aufweist und die darüber hinaus in allen Betriebsgrößenklassen im Mittel über dem Gesamtmittel liegt, bei allerdings auch stärkerer Streuung als in den übrigen Regionen. Eine gewisse Beziehung scheint also zwischen der Betriebsgröße und diesen (und weiteren, hier nicht gezeigten) Kennzahlen zu bestehen (womit natürlich nichts über eine Wirkungsrichtung gesagt ist).

⇒ Die Loess-Regressionslinien in den Abbildungen B4 bis B7 berücksichtigen neben der Betriebsgröße, die in Abbildung B4 über die Anzahl an Arbeitskräften konditioniert wird, zusätzlich eine Klassifizierung nach Jahren. Allerdings sind die beobachteten Trends in den drei betrachteten Jahren in fast allen Fällen sehr ähnlich, das heißt, obwohl sich möglicherweise absolute Werte von Jahr zu Jahr verändert haben mögen, bleiben die Variablenbeziehungen über die Jahre nahezu unverändert (dieser Frage wird noch detaillierter im folgenden Kapitel nachgegangen). Es wird aber noch eine Vielzahl anderer Aspekte sichtbar. So steht eine hohe Lohnquote in allen Jahren und allen Betriebsgrößenklassen in einem negativen Zusammenhang mit dem Betriebeseinkommen/AK (Abbildung B4); allerdings scheint der Effekt in sehr großen Betrieben weniger ausgeprägt zu sein als in kleineren. Der häufig zitierte und weiter unten noch diskutierte positive Zusammenahang zwischen Lohn je AK und Betriebseinkommen/AK, scheint demgegenüber vor allem in den Größenklassen 2 und 3 aufzutreten, während sehr kleine und sehr große Betriebe diesen Trend kaum zeigen. Bei der Flächenproduktivität (Betriebseinkommen/Eqm) wird hingegen ein weit weniger negativer Zusammenhang mit der Lohnquote beobachtet und auch die Beziehung zu Lohn je entlohnte AK ist wesentlich schwächer (Abbildung B5). Die Abbildungen in B6 verdeutlichen, daß zwischen Glasfläche/AK und der Arbeitsproduktivität auf der einen, und der Flächenproduktivität auf der anderen Seite, eher gegenläufige Trends zu bestehen scheinen. Während eine steigende Flächenleistung je AK mit einer Zunahme der Arbeitsproduktivität korrespondiert, geht mit ihr gleichzeitig eine Abnahme der Flächenproduktivität einher. Interessant sind auch die Beziehungen der Kennzahlen Spezialaufwand und Allgemeiner Aufwand zum Betriebseinkomen/AK. Steigender Spezialaufwand weist vor allem in den größeren Betrieben eine Beziehung zu abnehmenden Betriebseinkommen/AK auf, während die Beziehung steigender allgemeiner Aufwendungen vor allem in den kleineren Betrieben mit einem Rückgang der Arbeitsproduktivität korreliert ist.

∗ Die vier angesprochenen Abbildungen mit Loess-Regressionslinien stellen nur einen kleinen Teil aller möglichen Kombinationen von Kennzahlen und konditionierenden Variablen dar. Obwohl erneut und auch an dieser Stelle vor Überinterpretationen gewarnt werden muß (die Streuungen um die Linien sind erheblich) lassen sich doch durch die Trellis-Displays viele bemerkenswerte Gesichtspunkte herausarbeiten und darstellen. Wenngleich die hier vorgestellten Ergebnisse nicht durch ihre übergroße Eindeutigkeit bezüglich einer inhaltlichen Aussage bestechen, wird doch erneut unterstrichen, wie wertvoll die Darstellung von Beziehungszusammenhängen in Trellis-Displays ist.

3.2.2.2 Vergleich von Gruppen

Es wird nun die Frage untersucht, ob die unterschiedlichen, nach Region, Erhebungsjahr und Betriebsgröße[41] gebildeten Gruppen (insgesamt 24 Gruppen) durch ein gemeinsames Hauptkomponentenmodell beschrieben werden können (a-). Die Gruppenbildung erfolgt in Anlehnung an Gruppierungsmerkmale, die auch im Kennzahlenheft verwendet werden (Glasfläche und Jahr, AKBWL, 1996)) und unterscheidet die schon im vorigen Abschitt angesprochenen Regionen. Eine andere Gruppenbildung ist natürlich möglich und kann nach demselben, nun zu besprechenden Vorgehen, untersucht werden.

* Es soll zunächst geklärt werden, ob die Beziehungen zwischen den Variablen in den unterschiedlichen Gruppen annähernd gleich sind, oder ob zum Beispiel in großen Betrieben andere Wirkungszusammenhänge beobachtet werden als in kleineren Betrieben. Es wird demnach nicht untersucht, ob sich die Größe der einzelnen Werte in den Gruppen unterscheidet, ob also zum Beispiel der Rentabilitätskoeffizient über die Jahre zu- oder abgenommen hat. Vielmehr wird die Frage diskutiert, ob die Beziehung von, zum Beispiel Rentabilitätskoeffizient und Lohnquote, über die Jahre annähernd stabil ist oder nicht, und ob die Hauptquellen der Variabilität zum Beispiel in kleineren Betrieben andere sind als in größeren Betrieben. Gewählt wird für die Bearbeitung dieser diese Fragestellung nicht das Modell nach FLURY, 1988, da es zum einen die Multinormalverteilung und zum anderen die Verwendung der Kovarianzmatrix voraussetzt. Da beide Voraussetzungen hier nicht gegeben sind, wird auf die Modelle von KRZANOWSKI, 1979, und KERAMIDAS et al., 1987, zurückgegriffen, deren explorativer Charakter eher die Verwendung der standardisierten Werte und somit der Korrelationsmatrix in der Hauptkomponentenanalyse zulassen[42]. Die Hauptkomponentenanalyse beruht also immer auf der Korrelationsmatrix (die hier zur Erzielung der Gleichgewichtung aller Variablen eingesetzt werden muß, da die Skalen der Kennzahlen sehr unterschiedlich sind), und verwendet einen robusten Schätzer für die Objekte (nach CAMPBELL, 1980), das heißt, sehr extreme Werte gehen mit geringerem Gewicht in die Analyse ein als weniger auffällige Objekte.

Im Anschluß erfolgt eine kanonische Variablenanalyse, um mögliche Unterschiede der Mittelwertsvektoren der Gruppen graphisch sichtbar zu machen (b-).

a-Hauptkomponentenanalyse einzelner Gruppen

– Die Abbildungen B8 und B9 beinhalten für alle 24 Gruppen Plots mit den f_q-Werten nach

---

[41] Für die Betriebsgröße wird hier ein Shingle verwendet ohne einen Überlappungsbereich zu berücksichtigen, das heißt alle Betriebe werden eindeutig einer Betriebsgrößenklasse zugeordnet.

[42] Obwohl in den Orginalquellen ausschließlich mit der Kovarianzmatrix gearbeitet wird. Die Verwendung der Methoden bei Vorliegen standardiserter Werte wird in diesen abschließend nicht beantwortet. Allerdings ist sie sicher im explorativen Rahmen der vorliegenden Arbeit zulässig.

VELICER, 1976, beziehungsweise den W-Werten nach EASTMENT & KRZANOWSKI, 1982. Übersicht B12 gibt einige Informationen zu der Anzahl der Betriebe in den einzelnen Gruppen und den Gruppierungsfaktoren. Abbildung B10 zeigt Boxplots der Eigenwerte der 24 Gruppen.

⇒ Aus den Verfahren zur Ermittlung der Anzahl ‘wesentlicher’ Hauptkomponenten wird sichtbar, daß in den meisten Fällen eine, zwei oder drei Hauptkomponenten zur Repräsentation der Daten ausreichen. Das starke Gewicht der ersten beiden Hauptkomponenten zeigen auch die Boxplots in Abbildung B10. Die ersten Eigenwerte aller 24 Gruppen heben sich nicht nur im Mittel, sondern auch in den extremen Werten von der zweiten Hauptkomponente ab. Auch die Eigenwerte der zweiten Hauptkomponenten sind noch recht gut von den restlichen Komponenten abgegrenzt. Dann nehmen aber die Eigenwerte (und ihre Streuung) stark ab und die Überlappungsbereiche erheblich zu. Dies führt dazu, daß in der Betrachtung der Gamma-q-q-Plots nur bei den ersten und zweiten Eigenvektoren verläßliche Vergleiche mit den ‘typischen’ Eigenvektoren erstellt werden können.

– Die Abbildungen B11, B12 und B13 beinhalten diese Gamma-q-q-Plots für die ersten vier Eigenvektoren. Die Lage- und Skalenparameter der Gamma-Verteilung (geschätzt aus den Daten), sowie einige weitere Informationen zur Anpassung der Gamma-Verteilung an die quadrierten euklidischen Distanzen zwischen dem jeweiligen ‘typischen’ Eigenvektor und dem jeweiligen Eigenvektor der entsprechenden Gruppe, sind der Übersicht B13 zu entnehmen.

⇒ Vor allem beim ersten und zweiten Eigenvektor findet eine gute Anpassung der Distanzen an die Gamma-Verteilung bei Ermittlung der entsprechenden Parameter statt (Devianz < 1 bei 2 Freiheitsgraden). Die Residuen sind durchweg gering und nehmen erst beim dritten und vierten Eigenvektor zu. Die Gamma-q-q-Plots zeigen dementsprechend auch keine besonderen Auffälligkeiten. Allenfalls die Gruppen 6 und 21 fallen beim ersten Eigenvektor (Abbildung B11) stärker aus dem Rahmen. Bei den anderen abgebildeten Eigenvektoren gibt es verschiedentlich etwas vom Normalverlauf abweichende Gruppen, da aber die Trennung der Eigenwerte nur unscharf ist, und in keinem Fall eine besonders extreme Position auftritt, können keine ungewöhnlich auffälligen Gruppen identifiziert werden. Allerdings wird sichtbar, das die Distanzen mit jedem Eigenvektor zunehmen, das heißt beim ersten Eigenvektor liegt die mittlere Distanz noch bei 0,16, beim vierten Eigenvektor bereits bei 1,04 (Übersicht B13). Die Übereinstimmung zwischen den Gruppen wird also immer geringer, und nur in der ersten Dimension scheint es sich um eine allen Gruppen gemeinsame Komponente zu handeln (mit Ausnahme der Gruppen 6 und 21).

– Die Übersichten B14 und B15 beinhalten die Ergebnisse des Gruppenanalysemodells, die Abbildungen B14a und B14b und die Übersichten B16a und B16b die Hauptkomponentenergebnisse der auffälligen Gruppen 6, 7, 8 und 13.

∗ Während mit den Gamma-q-q-Plots einzelne Eigenvektoren miteinander verglichen werden,

vergleicht das Gruppenanalysemodell alle ausgewählten, in diesem Beispiel die ersten vier, Dimensionen, miteinander. Die delta-Werte in Übersicht B15 geben den Winkel an, in dem der jeweilige Eigenvektor der entsprechenden Gruppe zum mittleren Eigenvektor steht. Bei völliger Übereinstimmung ist delta demnach 0; stehen die Vektoren orthogonal zueinander nimmt delta den Wert 90 an. Die Summe der quadrierten Kosinusse von delta ist ein Maß für die Gesamtnähe der ausgewählten Dimension zur mittleren Konfiguration und hat ihr Maximum bei der Anzahl der Gruppen, das heißt je näher der Wert am Maximum liegt, desto näher sind die Gruppen der mittleren Konfiguration. Die mittleren Komponentenkoeffizienten definieren schließlich die mittlere Konfiguration, von der die Abweichungen bestimmt werden.

⇒ In der ersten Hauptkomponente wird eine recht große Übereinstimmung zwischen allen Gruppen festgestellt, Ausnahmen bilden vor allem die Gruppe 6 und die Gruppe 13. Besonders nahe am mittleren ersten Eigenvektor liegen die Gruppen 4 und 7. In der zweiten Dimension nimmt die Übereinstimmung insgesamt ab, die stärkste Abweichung zeigen nun die Gruppen 13 und 14, während die Gruppe 8 dem mittleren zweiten Eigenvektor am nächsten ist. In dritter und vierter Dimension findet dann in nahezu allen Gruppen eine recht große Abweichung vom Mittel statt. Es läßt sich also auch aus dieser Analyse auf eine oder zwei allen Gruppen gemeinsame Komponenten schließen, danach läßt die Übereinstimmung, die allerdings auch nur noch einen geringeren Anteil der Gesamtvarianz ‘erklärt’ deutlich nach. Insofern liefert diese Analyse eine ähnliche Einschätzung wie die Betrachtung der Gamma-q-q-Plots, obwohl natürlich hier und da Unterschiede zu beobachten sind, die auch darauf zurückzuführen sind, daß es sich beim Gruppenanalysemodell um die gleichzeitige und nicht, wie im Gamma-q-q-Plot, um die separate Analyse der ersten vier Dimensionen handelt.

⇒ Betrachtet man die mittleren Koeffizienten in Übersicht B14, so fällt bei der ersten Komponente die Trennung in negative Ladungen bei den Erfolgskriterien (beinkak, beinkeqm, beinkp, rdiffp, rentkoef) und positive Ladungen bei den Aufwandskriterien (allgawp, spez, lohnqp, fkp) auf mit Koeffizienten nahe am Nullpunkt bei den Strukturdaten (lohnak, heizqm, eqm, glasqm, glasqmak, anvermp). Die erste Dimension differenziert die Betriebe demnach nach Erfolg und Kosten, so daß erfolgreiche und kostensparende Betriebe auf der einen, und weniger erfolgreiche und kostenintensive Betriebe auf der anderen Seite stehen. Die positive Ladung bei kapkoef paßt genau in dieses Schema (siehe 3.2.2.1). In der zweiten Dimension sind es dann vor allem eqm, glasqm, glasqmak, anvermp und allgawp mit sehr negativen Koeffizienten und spezp, heizqm, lohnqp und fkp mit positiven Koeffizienten, so daß diese Dimension die Betriebe nach Fläche, Arbeitsintensität, Anlagevermögen und allgemeinem Aufwand auf der einen und Lohnquote, Energieverbrauch, Spezialaufwand und Fremdkapitaleinsatz auf der anderen Seite trennt, so daß man, wenn man denn will, von einer Technologie- und Vermögensdimension sprechen könnte. Interessant ist die Trennung der Erfolgskriterien in dieser Dimension in ebenfalls in zwei Gruppen, wobei beinkak, rdiffp und rentkoef mit der ersten Gruppe an Variablen korrespondieren (das heißt zum Beispiel geringer Rentabilitätskoeffizient bei geringer

Glasfläche/AK), während beinkp, beinkeqm und kapkoef eher mit der zweiten Gruppe korrelieren (das heißt zum Beispiel hohes Betriebseinkommen/Eqm bei hohem Energieverbrauch).

⇒ Abbildung B14 und Übersicht B16 dienen zur genaueren Betrachtung der besonders auffälligen Gruppen. Gruppe 7 ist dem mittleren Eigenvektor in der ersten Dimension (nach delta) am ähnlichsten, Gruppe 6 am unähnlichsten. Beim zweiten Eigenvektor sind die Gruppen 8 und 13 die dem mittleren Vektor am ähnlichsten beziehungsweise unähnlichsten Gruppen. Das CUSUM Diagramm zeigt für die 'typischen' Gruppen 7 und 8 die deutliche Beherrschung der ersten Komponente durch die Erfolgsvariablen, das heißt die erste Dimension wird vor allem durch die Streuung der erfolgreichen und weniger erfolgreichen Betriebe bestimmt, während sich die anderen Kriterien hier weit weniger finden. Wesentliche Unterscheidungsmerkmale der Betriebe sind demnach ihre Erfolgskennzahlen und weit weniger ihre Aufwands- Vermögens- oder Strukturdaten, so daß offensichtlich sehr erfolgreiche oder auch sehr erfolglose Betriebe ähnliche Werte bei anderen Kennzahlen besitzen können. Gruppe 6 unterscheidet sich hier sehr deutlich. Die Erfolgskennzahlen nehmen nur die Hälfte des ersten Eigenwertes in Anspruch, die anderen Kennzahlen liefern in fast demselben Umfang einen Beitrag zur Gesamtstreuung, insbesondere die Lohnquote. Darüberhinaus besitzen in dieser Gruppe eqm, glasqm, glasqmak und kapkoef stark negative Koeffizienten und zeigen somit in dieselbe Richtung wie die Erfolgsvariablen. In der zweiten Dimension fällt bei der auffälligen Gruppe 13 vor allem der der starke Beitrag der Flächenkennzahlen glasqm und eqm und die sehr geringe Bedeutung von lohnqp auf. Es läßt sich aber zusammenfassend festhalten, daß die Gruppierungen Glasfläche, Region und Erhebungsjahr in den wesentlichen ersten beiden Dimensionen eine erstaunliche Homogenität aufweisen, daß also das Beziehungsgefüge der hier ausgewählten Kennzahlen tatsächlich, von wenigen Ausnahmen abgesehen, stabil ist, das heißt, daß also in allen Gruppen die gleichen Korrelationen mehr oder minder Bestand haben. Der Frage, warum einzelne Gruppen von dieser allgemeinen Gültigkeit abweichen, wie die Gruppe 6 oder 13, soll in dieser Arbeit nicht weiter nachgegangen werden.

b-Kanonische Variablenanalyse

– Übersicht B17 enthält die Ergebnisse der kanonischen Variablenanalyse, die Eigenwerte der ersten beiden Dimensionen und die Mittelwerte der kanonischen Variablen. Abbildung B15 beinhaltet in B15a bis B15c die Lage der Mittelwerte der kanonischen Variablen, farblich gekennzeichnet nach Erhebungsjahr, Betriebsgröße und regionaler Lage. B15d bis B15f enthalten die konvexen Hüllen der Objektkonfigurationen in den ersten beiden Dimensionen für die Objekte aller 24 Gruppen (die Objekte selbst sind der Übersichtlichkeit halber nicht abgebildet), wobei die gleichen Farbkodierungen wie in B15a bis B15c gewählt werden.

⇒ Die ersten beiden kanonischen Variablen erfassen bereits über 90% der Gesamtvariablilität und zeigen eine sehr deutliche Trennung bei den kanonischen Variablenmittelwerten für die

Gruppierungsfaktoren Betriebsgröße (in der ersten Dimension) und regionale Lage (in der zweiten Dimension). Demgegenüber gibt es beim Gruppierungskriterium Erhebungsjahr erhebliche Überschneidungen, so daß sich die Mittelwertsvektoren der Jahre nur unwesentlich unterscheiden dürften. B15b legt darüber hinaus die Vermutung nahe, daß die Unterschiede zwischen den Regionen mit zunehmender Betriebsgröße abnehmen, zumindest bis zur dritten Glasflächenklasse (die Betriebsgröße nimmt von links nach rechts zu, die Farben entsprechen denen aus Übersicht B12). Allerdings ist die Streuung um die Mittelwerte nicht unerheblich. Dies wird durch die konvexen Hüllen um die Objektkonfigurationen verdeutlicht. Vor allem bei den Regionen gibt es nicht unwesentliche Überlappungsbereiche, während bei der Betriebsgröße eine insgesamt gesehen sehr gute Trennung zwischen den Gruppen erfolgt.

* Die konvexen Hüllen werden an Stelle von Konfidenzintervallen um die kanonischen Mittelwerte, wie sie in der Literatur häufig vorgeschlagen werden (siehe zum Beispiel KRZANOWSKI, 1988a) gewählt, um nicht implizit Annahmen zu treffen, vor allem hinsichtlich der Multinormalverteilung, die durch die Daten nicht gedeckt werden[43].

⇒ Die Abbildung B16 zeigt auf, daß es sich bei der Trennung der Betriebsgrößenklassen um eine tatsächlich in den Kennzahlen nachvollziehbare Gruppierung handelt. Die größten Betriebe weisen in der Mehrzahl der Kennzahlen auch die höchsten Werte auf, insbesonders bei Erfolgskennzahlen wie beinkak, rdiffp oder rentkoef, weisen aber gleichzeitig die niedrigsten Werte bei fkp und heizqm und niedrige Werte bei kapkoef, lohnqp und spezp auf. Die Betriebe der kleinsten Größenklasse zeigen genau das umgekehrte Verhalten. Eine Besonderheit stellt das Kriterium Betriebseinkommen/Eqm dar. Mit Zunahme der Betriebsgröße nimmt dieses offensichtlich ab, kleinere Betriebe haben also eine höhere Flächenproduktivität als größere Betriebe. Dies läßt den Schluß zu, daß das Betriebseinkommen bei der Flächenausweitung von Gartenbaubetrieben in geringerem Umfang zunimmt als die Betriebsgröße.

3.2.2.3 Gruppierung und Segmentierung

Während im vorangegangenen Kapitel im Mittelpunkt der Vergleich schon bestehender Gruppen behandelt worden ist, werden in diesem Abschnitt Methoden diskutiert, die zu einer Bildung von möglichst homogenen Gruppen führen können. Es kommen die Clusteranalyse (a-), Klassifikations- und Regressionsbäume (b-) und, als Sonderfall letzterer, der Chi Square Automatic Interaction Detector (CHAID) (c-) zum Einsatz.

a-Clusteranalyse

---

[43] Es ist grundsätzlich zu beachten, daß es sich bei der kanonischen Variablenanalyse nicht um ein annahmenfreies Verfahren handelt und homogene Kovarianzmatrizen und Multinormalverteilung vorausgesetzt werden. Insofern ist erneut der deskriptive Charakter der Analyse zu betonen.

Die Clusternanalyse stellt für diese Arbeit nur einen Randbereich dar, da es sich ja in erster Linie nicht um eine Methode zur Visualisierung von Daten, sondern zur gezielten Gruppierung von Objekten handelt. Es wird daher an dieser Stelle nur die Frage im Bereich der Clusteranalyse aufgegriffen, ob überhaupt von einer Clusterung der beobachteten beziehungsweise erhobenen Objekte ausgegangen werden kann.

* Es werden die modellbegründete Clusteranalyse, die nicht-hierarchische Klassifikation (Partitionierung um Medoide) und Fuzzy Clusterung, sowie die hierarchische agglomerative und divisive Vorgehensweise verwendet. Die Auswertung erfolgt getrennt nach Erhebungsjahren, um Unabhängigkeit der Objekte untereinander sicherzustellen. Für die modellbegründete Clusteranalyse werden die ausgewählten Kennzahlen (siehe Übersicht B18) standardisiert, für die anderen Verfahren erfolgt die Berechnung einer Proximitätsmatrix (nach KAUFMANN & ROUSSEEUW, 1990), die auf in Klassen unterteilten Variablenwerten beruht. Das heißt, alle Variablen werden in vier Klassen unterteilt, wobei die Klassengrenzen durch die Quartile gebildet werden. Zu Klasse 1 zählen demnach die Betriebe mit Werten bei der jeweiligen Variablen unterhalb des ersten Quartils, zu Klasse 2 Betriebe mit Werten zwischen dem ersten Quartil und unterhalb des Medians, zu Klasse 3 Betriebe mit Werten größer oder gleich dem Median bis zum dritten Quartil, und zu Klasse 4 schließlich Betriebe mit Werten oberhalb des dritten Quartils. Als modellbegründete Clusteralgorithmen werden verwendet: S, S*, Spherical, Unconstrained und Ward. Die Analysen werden unter den Annahmen des Zutreffens und des Nicht-Zutreffens der Multinormalverteilung durchgeführt. Im Bereich der nicht-hierarchischen Klassifikation erfolgt die Erstellung von Silhouettenplots für eine unterschiedliche Anzahl von Clustern. Die hierarchische Clusteranalyse, die auch auf der Proximitätsmatrix beruht, verwendet average-, complete- und single-link-Algorithmen im agglomerativen Teil, sowie ein divisives Vorgehen.

– Abbildung B17 beinhaltet die AWE-Werte der Clusteranaylsen der drei Erhebungsjahre für die unterschiedlichen Verfahren (in den Spalten) und nach dem normalen Vorgehen (in der unteren Zeile) und dem robusten Vorgehen (in der oberen Zeile). Die rote Referenzline ist bei 0 hinzugefügt. Es werden nur die AWE-Werte bis 20 Cluster gezeigt.

⇒ Bei Verwendung der robusten Methode gibt es in keinem Fall einen Hinweis auf das Vorliegen einer Gruppenstruktur, da alle AWE-Werte kleiner als 0 sind. Nur S und Ward zeigen bei Verwendung der nicht-robusten Methode einen leichten Ansatz möglicher Gruppierungen.

– Die Silhouettenplots in den Abbildungen B18a bis B18c basieren auf der Bildung von zwei bis neun Clustern nach der Methode der Partition um Medoide, die Abbildungen B19a bis B19c beinhalten die Silhouettenplots der Fuzzy Clusteranalyse für Lösungen mit zwei bis sechs Clustern.

⇒ Die Plots bestätigen, was auch schon durch die AWE-Werte angezeigt wird. Die mittlere

Silhouettenbreite in allen Fällen von unter 0,2 (Maximum bei völliger Einheitlichkeit aller Objekte eines Clusters ist 1), die zum Teil sogar negative Silhouettenbreite und der stark schräge Verlauf der Silhouetten sind Hinweise darauf, das in den gebildeten Clustern sehr heterogene Objekte zusammengefaßt sind, die sich nicht durch einige klare, gemeinsame Merkmalsausprägungen beschreiben lassen. Der niedrige Dunn-Koeffizient, der in allen Fällen '1/Anzahl der Cluster' entspricht, weist auf eine vollständige Fuzzy-Gruppierung hin, das heißt kein Objekt gehört eindeutig zu einem der gewählten Cluster.

– Die Abbildungen B20a bis B20d schließlich zeigen Dendrogramme und Bannerplots für die hierarchischen Clusterverfahren.

⇒ Die Aussagen der anderen Clusterverfahren werden erneut bestätigt. Die Dendrogramme besitzen Verschmelzungspunkte auf relativ hohem Niveau, deuten also ebenfalls auf heterogene Cluster hin, und auch die Bannerplots weisen eine geringe Trennschärfe zwischen den Gruppierungen auf, obwohl Unterschiede zwischen den Clusteralgorithmen auftreten. Die recht geringen agglomerativen Koeffizienten (zumindest bei single- und average-linkage) verdeutlichen, daß eine Vergrößerung der Cluster nur zu einer geringen Zunahme der Unähnlichkeiten in diesen Clustern führt, was wiederum ein Indiz für recht heterogene Gruppen ist.

∗ Die hier verwendeten Daten weisen nicht darauf hin, daß überhaupt eine Clusterstruktur vorliegt und eine Gruppierung daher mit einer entsprechenden Beliebigkeit verbunden ist. Diese Aussage bezieht sich allerdings nur auf die ausgewählten Variablen. Bei Auswahl anderer Kennzahlen können die Betriebe möglicherweise in sinnvolle Cluster getrennt werden, die zu einer guten Beschreibung der Betriebe, bezogen auf die betrachteten Merkmale, führen. Auf Grund der gerade gemachten Feststellungen ist jedoch nicht zu erwarten, daß sich diese Gruppierung, bezogen auf alle hier untersuchten Kennzahlen, fortsetzt. Die Diskussion um das Für und Wider von Clusteranalysen wird in Kapitel 4 noch einmal aufgegriffen. Die Verwendung der Proximitätsmatrix auf Grundlage der ordinalen Variablenklassen mag zunächst nicht ausreichend begründet erscheinen. Da diese Frage viele der nun folgenden Auswertungen und auch grundsätzliche Aspekte der Analyse der Kennzahlen betrifft, wird sie ebenfalls in Kapitel 4 noch einmal angesprochen.

b-Klassifikations- und Regressionsbäume (CART)

∗ Eine Alternative zur Clusteranalyse bieten die Klassifikations- und Regressionsbäume, die ebenfalls zu einer Einteilung einer großen Stichprobe in einzelne, möglichst homogene Segmente führen. Im Gegensatz zur Clusteranalyse sind aber die Entstehung der Segmente und deren wesentlichste Charakteristika durch die Baumstrukturen direkt nachvollziehbar und sichtbar. Allerdings richtet sich die Segmentierung an der Beziehung der Prediktorvariablen zu einer einzelnen abhängigen Variablen aus, und ist nicht, wie in der Clusteranalyse, durch eine

gleichwertige Einbeziehung aller Variablen in die Berechnung einer Proximitätsmatrix geprägt. Die Prediktorvariablen sind Übersicht B18 zu entnehmen. Als Zielvariable wird der Rentabilitätskoeffizient gewählt. Es wird sowohl mit dem vollen Datensatz und den Gewichtungen der multivariaten Ausreißeranalyse (siehe Übersicht B19), als auch mit einem um die Extremwerte verkleinerten Datensatz gerechnet, um der starken Abweichung der Kennzahl Rentabilitätskoeffizient von der Normalverteilung zu begegnen (siehe dazu die Übersichten B20 und B21, sowie die Normal-q-q-Plots in Abbildung B21). Beim Aufbau des vollen Regressionsbaums gelten die folgenden Einstellungen. Knoten werden nur gespalten, wenn mindestens 10 Objekte am Knoten vorliegen, eine Variable wird erst gespalten, wenn das kleinere Segment 5 Objekte umfaßt, und der Aufbau des Baumes stoppt, wenn die Knotendevianz von 0,01 unterschritten wird. Der verkleinerte Regressionsbaum wird bei sieben Terminalknoten betrachtet. Die Auswahl dieser Konstruktion ist eher willkürlich. Es werden jedoch bereits an dieser Stelle einige wichtige Gesichtspunkte erkennbar, die für die Beschreibung der Zusammenhänge zwischen abhängiger und Prediktorvariablen ausreichen sollen. Ein formaler Test auf die Güte dieser Lösung findet jedoch nicht statt. Da im Zentrum der Auswertung allerdings auch nicht die Formulierung eines speziellen Modells, sondern die Identifikation der stärksten Variablenbeziehungen steht, soll diese Vorgehensweise ausreichen.

– Die Abbildungen B22 bis B24 beziehungsweise B25 bis B27 beinhalten zusammenfassende Darstellungen der Ergebnisse der Analyse des vollen, gewichteten Datensatzes beziehungsweise des verkleinerten Datensatzes für die drei Erhebungsjahre getrennt. Sie zeigen den vollen Regressionsbaum (voll im Sinne der oben genannten Einstellungen), den auf sieben Terminalknoten gestutzten Baum mit vierklassigen Balkendiagrammen der abhängigen Variablen, die mittlere Residuendevianz verschiedener Lösungen im Prozeß des cost-complexity pruning, die mittlere Residuendevianz der sieben Terminalknoten, sowie zwei Residuenplots. Die Übersichten B22 bis B27 enthalten Informationen zu den betrachteten Modellen und den in den Regressionsbäumen auftauchenden Variablen.

⇒ Die Ergebnisse beider Analysen aller drei Jahre weisen große Ähnlichkeiten auf. In der obersten Ebene tritt immer die Lohnquote auf, darunter folgen Spezialaufwand und allgemeiner Aufwand. Die Aufwandsvariablen sind somit überwiegend an einer Segmentierung, ausgerichtet am Rentabilitätskoeffizienten, beteiligt. Andere Variablen kommen erst an weiter unten angesiedelten Knoten zur Nennung, am häufigsten (in allen Modellen) epertp, danach, in immerhin fünf Modellen anvermp und fremdakp. Regionale und Fremdkapital-Kennzahl tauchen in keinem Modell auf. Die Güte der Segmentierung, bezogen auf den Rentabilitätskoeffizienten, ist gut in den Barcharts zu erkennen (Teilabbildung b)), obwohl es natürlich einige Überschneidungen zwischen den Segmenten gibt. Die Terminalknotendevianzen (Teilabbildung d)) und die Residuen (Teilabbildung e)) sind insgesamt gering, die Residuen in der Regel annähernd normal verteilt (Teilabbildung f)), vor allem im verkleinerten Datensatz. Der Verlauf der Residuendevianzkurve (Teilabbildung c)) zeigt in allen Fällen, daß mit sieben

Terminalknoten schon deutlich über die Hälfte der Gesamtvarianz betrachtet werden kann. Werden die einzelnen Jahre (jetzt ausschließlich für die Lösung des verkleinerten Datensatzes) noch etwas genauer betrachtet, so fällt folgendes auf. 1992 (Abbildung B25) liegt der Schnittpunkt bei der Lohnquote bei 32,9%, das heißt praktisch am Median. Betriebe unterhalb dieses Wertes erreichen im Durchschnitt (bei Betrachtung der geschätzten Segmentwerte) immer höhere Rentabilitätskoeffizienten, als Betriebe, die den Median mehr oder weniger deutlich überschreiten. Nur bei einem Spezialaufwand deutlich oberhalb des Median (> 37,7%, Median 33,7%) oder des allgemeinem Aufwandes deutlich oberhalb des Median (> 30,2%, Median 25,5%) bleibt der Rentabilitätskoeffizient unter 1. Werden 32,9% bei der Lohnquote überschritten, wird selbst bei unterdurchschnittlichem Spezialaufwand im Mittel kein Rentabilitätskoeffizient von über 1 erzielt. 1993 (Abbildung B26) liefert annäherd dasselbe Bild. Der Schnittpunkt bei der Lohnqote liegt wieder bei 32,9%. Allerdings erzielen Betriebe mit recht hohem Spezialaufwand von über 37,8% (das ist ein Wert sehr nahe am dritten Quartil), selbst bei niedriger Lohnquote nur knapp über 0,9 liegende Rentabilitätskoeffizienten (allerdings recht breite Streuung in diesem Segment, siehe Barcharts in B26b). Überschreiten der 32,9%-Marke bei der Lohnquote führt aber auch in 1993 immer zu mittleren Rentabilitätskoeffizienten von unter 1. 1994 (Abbildung B27) zeigt dagegen, daß selbst bei recht hoher Lohnquote (im Bereich von 31,1% bis 44,1%), überdurchschnittlichem allgemeinen Aufwand (< 29,8%, Median 25,9%) und sehr geringen Spezielaufwand (< 27,5%, Median 31,5%), immerhin noch ein Segment mit Betrieben existiert, das im Durchschnitt einen Rentabiltätskoeffizienten von > 1 aufweist. Zusammenfassend läßt sich also festhalten, daß die drei Aufwandskennzahlen, die als Absolutwerte direkt in die Berechnung des Rentabilitätskoeffizienten eingehen, eine erwartungsgemäß starke Wirkung auf die am Rentabilitätskoeffizienten ausgerichtete Segmentierung besitzen. In gewissem Sinne bestätigt sich damit auch die Beobachtung aus der gruppenweisen Hauptkomponentenanalyse, die ja auch die erste und somit bestimmende Dimension in dem Gegesatz aus Erfolgs- und Aufwandskennzahlen ermittelt hat (siehe 3.2.2.2). Die - enttäuschend geringe - Beziehung zwischen den übrigen Kennzahlen und dem wichtigen Erfolgsmaßstab Rentabilitätskoeffizient kann ebenfalls erneut festgehalten werden.

c-CHAID

* Der Chi-Square Automatic Interaction Detector operiert nur mit kategorialen Daten. Verwendet werden daher die bereits unter a- beschriebenen, ordinalen Variablenklassen (ausgewählte Kennzahlen siehe Übersicht B18). Die Quartile der ausgewählten Kennzahlen, die die Klassengrenzen definieren, sind in Übersicht B28 festgehalten. Diese Kennzahlen werden nach Diskretisierung als ordinale Prediktorvariablen mit monotoner Kombinierbarkeit der Klassen verwendet (das heißt eine Zusammenlegung kann nur bei nebeneinanderliegenden Klassen erfolgen). Zusätzlich werden die Gruppierungsdaten Region und Absatzweg als nominale Variablen mit einer beliebigen Kombinierbarkeit der Klassen in die Analyse aufgenommen. Die Aufwandsvariablen allgemeiner Aufwand, Spezialaufwand und Lohnquote, die die CART-Lösung

so stark dominieren, bleiben nun unberücksichtigt, um die Struktur hinter diesen Kennzahlen näher zu betrachten. Als abhängige Variable wird wiederum der Rentabilitätskoeffizient gewählt. Es wird nach der ordinalen Methode gerechnet. Den Klassen des Rentabilitätskoeffizienten werden die Koeffizientenwerte der 12,5, 37,5, 62,5 und 87,5 Perzentile der einzelnen Jahre als Werte der jeweiligen Klassenmitte zugeordnet. Es findet die Bonferroni Anpassung in den Chi-Quadrat Tests statt. Ein neuer Split und die Verschmelzung von Kategorien erfolgt bei Unter- beziehungsweise Überschreiten der Signifikanzschwelle von 0,05. Als kleinstes, noch zu trennendes Segment werden 10 Objekte vorausgesetzt. Als kleinste Endsegmentgröße werden fünf Objekte festgelegt.

- Die Abbildungen B28 bis B30 beinhalten die Klassifikationsbäume mit den Schätzwerten für die abhängige Variable und den Prozentwerten der Prediktorvariablen in den einzelnen Klassen für die Auswertung der Jahre 1992 bis 1994. Die Übersicht B29 gibt Hinweise zu den gebildeten Segmenten und ihren Werten bei der abhängigen Variablen. In den Abbildungen B31 bis B33 sind für die wichtigsten Prediktorvariablen Balkendigramme und ein Rugplot der abhängigen Variablen, gruppiert nach den Segmenten auf der untersten hier betrachteten Ebene, dargestellt.

⇒ Als am stärksten trennende Kennzahl tritt 1992 und 1994 die Betriebsgröße über die Kennzahl Eqm auf. In der zweiten Ebene folgen dann in diesen beiden Jahren die Region und 1992 der Anteil Fremd-AK beziehungsweise 1994 das Fremdkapital. 1993 stehen Fremdkapital und Region in den ersten beiden Ebenen. Auf der dritten und vierten Ebene werden sich die Lösungen der einzelnen Jahre wieder unähnlicher, 1992 tauchen glasqmak, netinvp und anvermp auf, 1993 ak, fremdakp, lohnak und verm, und 1994 heizqm, netinvp, und lohnak. Interessant ist aber, daß alle drei Lösungen die regionale Lage in einer der beiden oberen Ebenen beinhalten, und daß auch die Betriebsgröße einen starken segmentierenden Effekt hat. Darüber hinaus tritt die Kennzahl Anteil Fremdkapital, die in der CART Lösung überhaupt nicht auftaucht, hier deutlich in den Vordergrund. Weitere Splits der Klassifikationsbäume sind an den in den Abbildungen angegebenen Stellen möglich, die Gruppen werden dann aber sehr klein. Durch die Baumstruktur lassen sich die entstandenen Segmente gut beschreiben. So handelt es ich 1992 bei den erfolgreichsten Betrieben mit einem geschätzten durchschnittlichen Rentabilitätskoeffizienten von 1,22 um überdurchschnittlich große Betriebe aus Region 1, die sehr stark investiert haben. 1994 liegen die besten Betriebe mit einem geschätzten durchschnittlichen Rentabilitätskoeffizienten von 1,26 in den mittleren Bertriebsgrößenklassen, mit unterdurchschnittlichem Fremdkapitalanteil und sehr geringem Energieverbrauch je Quadratmeter. Analog lassen sich die übrigen Segmente interpretieren. Allerdings ist kein einheitliches Schema in den drei Erhebungsjahren zu erkennen, das heißt die Variablen, die die Segmente am besten beschreiben unterscheiden sich von Jahr zu Jahr (obwohl es sich um dieselben Betriebe handeln kann). Die Balkendiagramme zeigen überzeugend, daß den Variablen in den ersten beiden Ebenen der Klassifikationsbäume tatsächlich sichtbare Effekte zugrunde liegen. In allen drei Jahren hat die Region 1 deutlich mehr Betriebe in den hohen

Rentabilitätsklassen als die übrigen Regionen. Die Zunahme des Erfolgs mit Zunahme der Betriebsgröße beziehungsweise mit Abnahme des Anteils Fremdkapital wird ebenso deutlich. Diese Betrachtung paarweise Beziehungen ist aber nicht unproblematisch, da ein beobachteter Effekt in derartigen Darstellungen möglicherweise nur Ergebnis der starken Korrelation beider Variablen mit einer dritten Variablen ist (Simpsons' Paradox, siehe 2.4.2). Schließlich sind die Rugplots in den Abbildungen B31 bis B33, die die Orginalwerte der Zielvariablen für die einzelnen Segmente enthalten, eine Mahnung zur Vorsicht in der Interpretation der Klassifikationsbäume. Zwar ist in allen Fällen ein leichter Trend sichtbar, und die Segmente scheinen sich hinsichtlich ihres Rentabilitätskoeffizienten etwas zu unterscheiden, es wird aber auch deutlich, daß die Überschneidungen enorm sind, und die Abgrenzung der Segmente gegeneinander nur recht undeutlich ist.

3.2.2.4 Diskrete graphische Modelle

Diskrete graphische Modelle werden nun eingesetzt, um die Beziehungen der Kennzahlen untereinander zu untersuchen. Die Kennzahlen werden wie für die im vorangegangenen Kapitel beschriebene CHAID-Analyse diskretisiert. Die Anwendung kontinuierlicher graphischer Modelle oder gemischter graphischer Modelle bietet sich aufgrund der starken Abweichungen von der Multinormalverteilung nicht an. Zunächst werden 15 Kennzahlen, darunter eine Erfolgskennzahl und 14 weitere Kennzahlen in die Untersuchung der Jahre 1992 bis 1994 einbezogen (a-). Anschließend erfolgt die Betrachtung der Beziehungen von sechs Erfolgskennzahlen untereinander (b-).

a-Erfolgs- und andere Kennzahlen

Für jedes Erhebungsjahr werden sechs graphische Modelle gesucht. Jedes Modell beinhaltet eine der sechs Erfolgskennzahlen Betriebseinkommen/AK (beinkak), Betriebseinkommen/Eqm (beinkeqm), Betriebseinkommen in % BE (beinkp), Kapitalkoeffizient (kapkoef), Reinertragsdifferenz in % BE (rdiffp), oder Rentabilitätskoeffizient (rentkoef) und 14 weitere Kennzahlen, und zwar diejenigen, die in Übersicht B18 auch für die CHAID-Analyse aufgelistet sind. Die Aufwandskennzahlen werden also auch in dieser Auswertung nicht berücksichtigt, da sich das Interesse auf die Beziehungen der Erfolgskennzahlen zu denjenigen Kennzahlen richtet, die nicht direkt in ihre Berechnung einfließen, sondern von denen eher eine strukturelle Wirkung erwartet wird, die sich möglicherweise in den Erfolgskennzahlen niederschlägt.

* Die Vielzahl der Kennzahlen führt zu sehr schwach besetzten Tabellen, die in sehr vielen Zellen eine 0 besitzen. Es werden daher exakte Tests mit 500 Simulationen zur Modellfindung eingesetzt. Das Signifikanzniveau für Einschluß oder Ausschluß einer Verbindung zwischen zwei Kennzahlen (das heißt für die Ablehnung oder Bestätigung der Hypothese bedingter Unabhängigkeit) liegt bei p = 0,05. Als Suchalgorithmen werden die Rückwärts-Elimination ausgehend vom vollen Modell und die Methode nach EDWARDS & HAVRÁNEK, 1987, gewählt.

Bei letzterer erfolgt im ersten Schritt die Ermittlung des minimalen, unbestimmten Modells. Es wird dann geprüft, ob dieses Modell zur Beschreibung der Daten ausreicht, und welche Modelle durch Hinzufügen weiterer Verbindungen zum minimalen Modell ebenfalls akzeptable Modelle für die vorhandenen Daten liefern. So ergibt sich im letzten Schritt eine Liste möglicher Modelle, in denen manche Verbindungen immer, andere Verbindungen nur in einem Teil der möglichen Modelle fehlen. Das Verfahren verdeutlicht also die dem Modellierungsprozeß inhärente Unsicherheit. Es werden ausschließlich ungerichtete Graphen verwendet. Dies ist auch gerade im Blick auf die Kennzahlen sinnvoll, da die Wirkungsrichtung nicht eindeutig ist. Die Wechselwirkungen der Kennzahlen untereinander werden somit unterstrichen.

– Da nur Tabellen mit acht Variablen mit der dem Verfasser zur Verfügung stehenden Software (DIGRAM von Svend Kreiner) verrechnet werden können, erfolgt zunächst ein Screening der Variablen auf direkte Beziehungen in allen Zwei- und Drei-Wege-Tafeln. Abbildung B34 zeigt im oberen Teil das gesamte Beziehungsgeflecht im vollen Modell zum Beispiel für 1993 mit der Erfolgskennzahl Betriebseinkommen/AK. Im unteren Teil der Abbildung sind dann alle nach dem Screening ermittelten bedingt unabhängigen Verbindungen entfernt. Derselbe Prozeß wird für alle sechs Erfolgskennzahlen und alle drei Jahre wiederholt. Daraus resultiert Übersicht B30, die zeigt, zwischen welchen Variablen in den einzelnen Jahren direkte Beziehungen bestehen.

⇒ Die Ergebnisse des Screenings in den drei Erhebungsjahren sind sich, bezogen auf die Beziehungen der ‘übrigen’ Kennzahlen zu den Erfolgskennzahlen, relativ ähnlich. So sind zum Beispiel beim Betriebseinkommen/AK immer direkte Beziehungen von beinkak zu eqm, fkp, glasqmak und lohnak vorhanden. 1992 kommen noch netinvp und abswg, 1994 epertp hinzu. Eine vergleichbare Übereinstimmung ergibt sich zwischen den Jahren auch bei den anderen Erfolgskennzahlen, das heißt, daß von einer gewissen Kontinuität der Beziehungen in den Erhebungsjahren ausgegangen werden kann. Die Variablen, zu denen direkte Beziehungen bestehen unterscheiden sich demgegenüber bisweilen auffällig von Kennzahl zu Kennzahl. Nur der Fremdkapitalanteil taucht in allen Modellen, mit einer Ausnahme (kapkoef, 1992), auf. Weitere, häufig auftauchende Kennzahlen sind eqm (in 12 Modellen), lohnak (in 11 Modellen), heizqm (in 10 Modellen) und glasqmak (in 9 Modellen). anvermp taucht nur in drei Modellen, dafür aber konstant in denen mit der Kennzahl Kapitalkoeffizient auf. glasqm und verm sind in keinem der 18 Modelle genannt. Jede Erfolgskennzahl ist somit mindestens teilweise ein Spiegel von Informationen, die von den anderen Erfolgskennzahlen nicht wiedergegeben werden.

– Für 1993 werden die Kennzahlen mit direkten Beziehungen nun der intensiven Analyse mit exakten Tests unterzogen. Die verwendeten Kennzahlen, sowie die Ergebnisse der Rückwärts-Elimination und der EH-Prozedur sind in Übersicht B31 festgehalten. Es ergeben sich nach Rückwärts-Elimination die sechs graphischen Modelle in Abbildung B35.

⇒ Die Modelle sind durch eine Vielzahl von Variablenbeziehungen gekennzeichnet. Die direkten

Beziehungen, die im Screening beobachtet werden, sind auch im reduzierten Variablensatz nach wie vor vorhanden. Bedingte Unabhängigkeiten, vor allem mit Blick auf die Erfolgskennzahlen, sind praktisch nicht vorhanden. Die Modelle werden häufig durch Cliquen beherrscht, also maximal komplette Subgraphen (ein maximal kompletter Subgraph ist dadurch gekennzeichnet, daß alle Variablenpunkte einander benachbart sind). Zwischen den durch das Screening ausgewählten Variablen bestehen vielfältige direkte und auch indirekte Beziehungen. Hinter den Variablen. die in den Modellen abgebildet sind, stehen alle übrigen Variablen, die entweder keine besonders starke direkte Beziehung zu einer der Erfolgskennzahlen besitzen oder bedingt unabhängig von ihr sind, das heißt unabhängig von ihnen, gegeben die im Modell berücksichtigten Variablen.

⇒ Das graphische Modell mit der Arbeitsproduktivität (beinkak, B35a) ist gekennzeichnet durch eine die vier Kennzahlen beinkak, eqm, glasqm und lohnak umfassende Beziehung, sowie eine von den übrigen Kennzahlen (gegeben beinkak) unabhängige Wechselwirkung mit fkp. Das Modell für die Flächenproduktivität (beinkeqm, B35b) bestätigt die direkten Beziehungen zu ak, lohnak, glasqmak, heizqm und fkp aus dem Screening, und enthält Querverbindungern aller Kennzahlen zu mindestens zwei weiteren Variablen. Etwas stärker vereinfacht wirkt das Modell für die Wertschöpfungsquote (beinkp, B35c). Eine Clique von beinkp, lohnak und fkp wird ergänzt durch eine, gegeben beinkp, von lohnak und fkp unabhängige, Wechselwirkung zwischen heizqm und beinkp. Der wohl am stärksten verbundene Graph ist der für das Modell, das den Kapitalkoeffizienten beinhaltet (B35d). Die Bildung von vier maximal vollständigen Subgraphen ist möglich und verdeutlicht das enge Geflecht von anverp, fkp, lohnak und fremdakp, in dessen Mittelpunkt immer der Kapitalkoeffizient steht. Reinertragsdifferenz und Rentabilitätskoeffizient liefern (B35e und B35f) einander recht ähnliche Modelle, mit direkten Beziehungen zwischen den Erfolgskennzahlen und region, fkp und fremdakp. Interessanterweise taucht in beiden Modellen auch die Kennzahl eqm auf, die aber gegeben die anderen Variablen von den Erfolgskennzahlen unabhängig ist, der Betriebsgrößeneffekt also bei diesen beiden Kennzahlen durch andere Variablen bereits erklärt wird.

– Die Abbildungen B36 bis B40 fassen einige, nicht alle, der, im wesentlichen direkten, Beziehungen in Form von Trellis-Displays zusammen. In den Panels, die in der Regel in den drei Spalten durch die drei Erhebungsjahre und in den Zeilen durch die in vier Klassen gruppierten Erfolgskennzahlen konditioniert sind, befinden sich die relativen Häufigkeiten der vier Klassen der im Einzelfall betrachteten Kennzahl.

⇒ Die Abbildungen B36a und B36b zeigen den Zusammenhang zwischen Betriebsgröße (eqm), ‘Fläche zu Arbeit’-Verhältnis (glasqmak) und Arbeitsproduktivität (beinkak). Größere Betriebe gehören in der Mehrzahl der Fälle zu den Betrieben mit höherer Arbeitsproduktivität. Das gleiche gilt für Betriebe mit höherem ‘Fläche zu Arbeit’-Verhältnis (glasqmak). Die Flächenproduktivität (beinkeqm) hat nahezu die entgegengesetzten Beziehungsrichtungen (Abbildungen B36c und

B36d), das heißt kleinere Betriebe besitzen eher eine hohe Flächenproduktivität als größere Betriebe, und ein weites ‘Fläche zu Arbeit’-Verhältnis korrespondiert häufiger mit einer geringen als mit einer hohen Flächenproduktivität. Es bestätigt sich damit eine Beobachtung aus der kanonischen Variablenanalyse (siehe 3.2.2.2).

– Das Beziehungsgeflecht von Betriebsgröße und ‘Fläche zu Arbeit’-Verhältnis soll noch näher betrachtet werden. Im Modell für die Flächenproduktivität taucht die Kennzahl Einheitsquadrameter nicht auf, sie wird im Screening als bedingt unabhängig, gegeben (unter anderem) Anzahl AK, identifiziert. Deswegen wird in der Folge auch die Anzahl AK als Maß der Betriebsgröße mitverwendet.

⇒ Die Abbildungen B37a und B37b zeigen die zu erwartenden Effekte, und zwar, daß das ‘Fläche zu Arbeit’-Verhältnis mit zunehmender Fläche zunimmt und mit zunehmender Anzahl AK abnimmt. Auf der anderen Seite zeigt B37c das auch Anzahl AK und Einheitsquadratmeter in enger Beziehung stehen und beide Kennzahlen in gewissem Sinne ein Maß der Betriebsgröße sind, Betriebe mit einer großen Fläche also auch überwiegend viele Arbeitskräfte und Betriebe mit vielen Arbeitskräften auch große Flächen haben. Einheitsquadratmeter, Anzahl AK und ‘Fläche zu Arbeit’-Verhältnis sind in B37d nur für das Jahr 1993 noch einmal abgebildet. Die vorangegangenen Beobachtungen werden zwar bestätigt, aber es fällt doch auf, daß die Zunahme des ‘Fläche zu Arbeit’-Verhältnisses durch Steigerung der Betriebsgröße auf den einzelnen Stufen der Anzahl AK nicht gleich verläuft.

– Dieser Zusammenhang wird in anderer Weise auch noch einmal in B38 aufgegriffen. Die Panels beinhalten hier nun aber wieder Loess-Regressionslinien (in ihrer Definition identisch zu denen aus 3.2.2.1), es werden die log-transformierten Ausgangsvariablen verwendet und die Klassenbildung erfolgt nach einem equal-count-Algorithmus mit einem großen Überlappungsbereich.

⇒ Am auffälligsten ist B38a. In der Gruppe der kleinsten Betriebe stehen Steigerung der Fläche und Zuwachs des ‘Fläche zu Arbeit’-Verhältnisses in einem engen Zusammenhang. Dann scheint jedoch eine Art Plateau erreicht zu werden (um das herum natürlich auch hier eine nicht unerhebliche Streuung zu verzeichnen ist). Offensichtlich gibt es also einen Wert beim ‘Fläche zu Arbeit’-Verhältnis, über den hinaus selbst recht große, am Kennzahlenvergleich beteiligte Betriebe, im Mittel nicht hinauskommen und der auch schon von den durchschnittlich großen Betrieben erreicht wird. Dieser Wert liegt bei log(7,5) also ungefähr 1100 qm/AK, und damit praktisch am Median. Bei Konditionierung in den Spalten nach log(ak) in B38d scheint sich dagegen der abwärts gerichtete Trend des ‘Fläche zu Arbeit’-Verhältnises in allen AK-Klassen sehr ähnlich fortzusetzen (mit Ausnahme von Jahr 1992). B38b und B38c zeigen im Grunde dieselben Effekte wie B37d. Interessant ist aber vor allem in B38b der ein wenig flacher werdende Verlauf der Regressionslinien mit zunehmender Betriebsgröße, das heißt, daß eine Steigerung in der Anzahl der AK in den größeren Betriebsgrößenklassen weniger negativ mit

dem ‘Fläche zu Arbeit’-Verhältnis in Beziehung steht als in den kleineren Betriebsgrößenklassen. Erneut wird sichtbar, daß mit zunehmender Betriebsgröße auch ein Anstieg des ‘Fläche zu Arbeit’-Verhältnisses einhergeht.

– Abbildung B39 beinhaltet zwei Trellis-Displays zum Modell des Kapitalkoeffizienten, die zwei interessante Beobachtungen zulassen.

⇒ Zum einen haben offensichtlich Betriebe mit einem geringen Anlagevermögen in % des Vermögens einen im Durchschnitt niedrigeren Kapitalkoeffizienten als Betriebe mit einem höheren Anlagevermögen (B39a). Wird anvermp als ein Kriterium für die Beurteilung der Modernität von Betrieben verstanden, so mag diese Beobachtung als ein insgesamt negativer Technologieeffekt interpretiert werden, das heißt eine sehr starke Modernisierung (charakterisiert durch hohe anvermp-Werte) korrespondiert mit hohen Kapitalkoeffizienten und somit mit einer eher unbefriedigerenden Kapitalausnutzung. Ein positiver Technologieeffekt würde sich in fallenden Kapitalkoeffizienten bei steigendem Anlagevermögen auszeichnen. Es mag aber auch sein, daß hier der Kapitalkoeffizient kein besonders gut geeigneter Maßstab ist, da vom Betriebseinkommen noch der Lohnaufwand für Fremd- und und der Lohnanspruch der Familien-AK abgedeckt werden muß und möglicherweise überhaupt erst beim Lohnaufwand die Modernisierungseffekte besonders deutlich hervortreten. Die zweite Beobachtung betrifft das Fremdkapital (B39b). Hier ist es so, daß Betriebe mit niedrigen Kapitalkoeffizienten, das heißt einer effizienten Kapitalausnutzung, in ihrer Mehrzahl entweder sehr geringe oder sehr hohe fkp-Werte haben, während Betriebe mit einer weniger guten Kapitaleffizienz (kapkoef Klassen 3 und 4) beim Fremdkapitaleinsatz eher im mittleren Bereich liegen.

⇒ B39c und B39d zeigen exemplarisch für Reinertragsdifferenz und Rentabilitätskoeffizient, die sich auch bei allen anderen Erfolgskennzahlen, außer dem Kapitalkoeffizienten, wiederholende Beobachtung, daß Betriebe im Bereich hoher Erfolgskennzahlen eher niedrige fkp-Werte haben, und weniger erfolgreiche Betriebe hohe fkp-Werte aufweisen.

– B40 schließlich beschäftigt sich mit den Regionsunterschieden, die vor allem bei Reinertragsdifferenz (B40a) und Rentabilitätskoeffizient (B40b) zu Tage treten.

⇒ Der Anteil der Betriebe aus Region 1 (das entspricht in den Trellis-Displays der Klasse 2) liegt bei den erfolgreicheren Betrieben immer über dem Anteil der Betriebe aus den übrigen Regionen (hier Klasse 1), bei den weniger erfolgreichen Betrieben ist es fast immer umgekehrt.

– B40c und B40d verdeutlichen eine - wenn auch nicht sehr gut sichtbare - bedingte Unabhängigkeit.

⇒ Aus dem Modell für das Betriebseinkommen/AK und dem Screening wird ersichtlich, daß das Betriebseinkommen/AK und Region unabhängig voneinander sind, gegeben glasqmak. B40c bestätigt, daß es zwar Regionsunterschiede gibt, daß diese aber in den vier glasqmak-Klassen (in den Spalten) einander sehr ähnlich sind. B40d zeigt dagegen dieselbe Konstellation bezogen

auf den Rentabilitätskoeffizienten, für den (rentkoef) unabhängig (glasqmak) gegeben (region) gilt. Beide Regionen weisen deutlich voneinander abweichende Verläufe bei glasqmak auf, innerhalb der Regionen sind sie jedoch in allen Rentabilitätskoeffizinetenklassen relativ konstant (auffällige Ausnahme rentkoef-Klasse 4 in sonstigen Regionen (Klasse 1)).

b-Erfolgskennzahlen

- Es werden nun die sechs Erfolgskennzahlen gemeinsam in ein graphisches Modell einbezogen, um die Beziehungen der Erfolgsvariablen untereinander näher zu betrachten. Übersicht B32 beinhaltet die Ergebnisse des Modellfindungsprozesses nach Rückwärts-Elimination und EH-Prozedur, Abbildung B41 die graphischen Modelle für die drei Erhebungsjahre nach der Rückwärts-Elimination. Die Ergebnisse der Rückwärts-Elimination werden hier, wie unter a-, gewählt, da sie erneut die etwas einfacheren, mit den zu Daten vereinbaren Modelle liefern und im Vergleich mit den Ergebnissen der EH-Prozedur zu einem sehr ähnlichen Ausschluß von Variablenverbindungen und damit Aussagen zur bedingten Unabhängigkeit kommen.

⇒ Allerdings wird die recht schwache Verbindung zwischen beinkeqm und beinkp nach EH-Prozedur 1993 und 1994 aus allen akzeptablen Modellen und 1992 aus drei der sechs akzeptablen Modelle ausgeschlossen, während diese Beziehung in den Modellen nach Rückwärts-Elimination für 1993 und 1994 erhalten bleibt. Die drei graphischen Modelle weisen recht große Unterschiede auf, und eine allgemeine Gesetzmäßigkeit läßt sich aus diesen drei Jahren nicht ableiten. Die starken Beziehungen zwischen den Erfolgskennzahlen werden vor allem durch den Graph für das Erhebungsjahr 1992 verdeutlicht. 1993 und 1994 ergeben sich ein wenig differenziertere Bilder. Vor allem in 1993 deutet sich eine Trennung in zwei Kennzahlengruppen an, bestehend aus beinkeqm, beinkp und kapkoef auf der einen und rentkoef, rdiffp und beinkak auf der anderen Seite, die über die Beziehung von Wertschöpfungsquote (beinkp) und Rentabilitätskoeffizient beziehungsweise Kapitalkoeffizient und Reinertragsdifferenz in Verbindung stehen. Abbildung B42 betrachtet das obere, Abbildung B43 das untere Dreieck dieses Modells. Flächenproduktivität und Wertschöpfungsquote zeigen sichtbare negative Korrelationen mit dem Kapitalkoeffizienten, während beinkp als einzige der Erfolgskennzahlen eine einigermaßen deutlich positive Korrelation zu beinkeqm aufweist und somit als Bindeglied zum Rentabilitätskoeffizienten verständlich wird. Unter den Kennzahlenbeziehungen in Abbildung B42 fällt vor allem die fast vollständige Korrelation von Reinertragsdifferenz und Rentabilitätskoeffizient, sowie die sehr starke Beziehung von Arbeitsproduktivität und Rentabilitätskoeffizient auf. Diese drei Kennzahlen führen demnach zu vergleichbaren Einstufungen der Betriebe. Die Beziehungen zu Wertschöpfungsquote und Kapitalkoeffizient sind ebenfalls vorhanden, aber deutlich geringer. Bemerkenswert ist jedoch, daß auch hier, wie schon an anderen Stellen in dieser Arbeit deutlich wird, daß unterschiedliche Erfolgskennzahlen zu unterschiedlichen Beurteilungen und Aussagen zum Erfolg der Betriebe führen. Wenn die Betriebe mit Hilfe von Erfolgskennzahlen in Gruppen eingeteilt werden sollen,

wie es zum Beispiel in der Einteilung in erstes Drittel und drittes Drittel im Kennzahlenheft des Arbeitskreises Betriebswirtschaft geschieht, ist zu beachten, daß es vor allem in den mittleren Bereichen der Kennzahlenwerte bei Betrieben mit ähnlichen Erfolgskennzahlen erhebliche Unterschiede bei den übrigen Kennzahlen gibt. Eine, an welcher einzelnen Kennzahl auch immer, festgemachte Gruppierung der Kennzahlenbetriebe kann immer nur zu einer Teilbetrachtung des Erfolges eines Betriebes dienen. Andererseits kann demzufolge eine Gruppierung aufgrund eines Kennzahlenmixes (siehe folgender Abschnitt 3.2.2.5) auch letztlich nur zu einer ähnlich willkürlichen Gruppierung dienen, zumal bereits mehrfach festgehalten worden ist, daß die Beziehungen der Erfolgskennzahlen zu anderen Kennzahlen zum Teil erheblich voneinander abweichen.

3.2.2.5 Formale Begriffsanalyse

Die hierarchischen Liniendiagramme der formalen Begriffsanalyse unterstützen die interaktive Exploration umfangreicher Datensätze. Es handelt sich bei der formalen Begriffsanalyse, so wie sie hier eingesetzt wird, weniger um eine Analysetechnik als vielmehr um ein Instrument der Erkundung großer Datenmengen. Der Einsatz der Liniendiagramme, der in Schriftform allerdings nur unbefriedigend dargestellt werden kann, wird an mehreren Beispielen demonstriert. Im ersten Teil geht es um die Beantwortung konkreter Fragestellungen, die im Zusammenhang mit den Beziehungen der Kennzahlen untereinander auftauchen (a-). Im zweiten Teil erfolgt beispielhaft die Identifikation interessanter Betriebsgruppen durch hierarchische Liniendiagramme (b-).

a-Konkrete Fragestellungen

* Die formale Begriffsanalyse operiert nur mit diskreten Daten. Kontinuierliche Daten, wie sie im Kennzahlenvergleich vorliegen, müssen vor Verwendung in den Liniendiagrammen also entsprechend transformiert beziehungsweise, um in der Sprache der Begriffsanalyse zu bleiben, begrifflich skaliert werden. Die begriffliche Skalierung bietet eine Vielzahl von Möglichkeiten, Vorwissen in die Klassenbildung und somit in das Aussehen der Liniendiagramme mit einzubringen. Da im Zusammenhang mit den Kennzahlen allerdings nur wenige Werte eine inhaltlich vorgegebene Bedeutung haben - ein Beispiel ist der Rentabilitätskoeffizient mit einer klaren Schnittstelle bei 1 - und vielfach nur der Vergleich mit anderen Betrieben ein Maßstab dafür ist, ob ein Wert hoch oder niedrig ist, wird die Klassenbildung, die auch für CHAID und die diskreten, graphischen Modelle verwendet wird, im wesentlichen beibehalten. Bisweilen wird sie durch inhaltlich sinnvolle Klassenbildungen ergänzt. Als Skalierungstyp wird in den meisten Fällen eine Biordinalskala eingesetzt, das heißt, das Liniendiagramm teilt sich in zwei Stränge auf, die ihrerseits von unten nach oben sortiert sind. Praktisch bedeutet dies, daß die meisten Liniendiagramme sechs Begriffe umfassen, von denen vier von besonderer Bedeutung sind. Unten links stehen die Betriebe mit Werten unterhalb des ersten Quartils; oben links die Betriebe mit Werten zwischen erstem Quartil und Median; oben rechts die Betriebe mit Werten zwischen dem Median und dem dritten Quartil und unten rechts Betriebe mit Werten oberhalb

des dritten Quartils. In der Regel werden, der Übersichtlichkeit halber, nicht mehr als zwei Liniendiagramme ineinander verschachtelt. Das äußere Liniendiagramm steht dann in der Beschriftung der Abbildung (oben links) oben, das innere Liniendiagramm unten. Die Beschriftung der Begriffe erfolgt in Prozentangaben der Betriebe in den einzelnen Begriffen und unter Einbeziehung aller Objekte im Begriff, so daß, entsprechend der Leseregel bei Liniendiagrammen, zum Beispiel oben rechts der Prozentanteil der Betriebe steht, die bei der betrachteten Kennzahl Werte haben, die größer oder gleich dem Median sind (bis zum Maximum), während unten rechts der Anteil der Betrieb steht, die Werte oberhalb des dritten Quartils haben. Die linke Seite des Liniendiagramms stellt die Betriebe unterhalb des Medians, die rechte Seite die Betriebe oberhalb des Medians dar, und die zwei Begriffe auf jeder Seite des Liniendiagramms differenzieren diese Unterteilung noch weiter. Es handelt sich also im Grunde um eine ungewöhnlich aufgebaute Zwei-Wege-Tafel. Die Prozentwerte basieren auf den Werten aller drei Erhebungsjahre gemeinsam.

- Die Liniendiagramme werden nun eingesetzt, um die Daten zu konkreten Fragestellungen sichtbar zu machen. Als Fragen werden einige der Feststellungen (Hypothesen oder Schlußfolgerungen) aus der Arbeit von BITSCH, 1994, verwendet, und es ist zu prüfen, ob die Aussagen mit den Werten in den Liniendiagrammen vereinbar sind oder nicht.

- Feststellung 1: „Mit dem Betriebserfolg steigt die Entlohnung der Arbeitskräfte an“ (BITSCH, 1994, Seite 156). Diese Feststellung wird von BITSCH, 1994, tendenziell bestätigt; vor allem die sogenannte Spitzengruppe besonders erfolgreicher Betriebe[44], weist eine höhere Entlohnung je AK auf als der Rest der Betriebe. Die Abbildungen B44 bis B48 zeigen in Liniendiagrammen die Beziehungen zwischen Lohn je entlohnte AK zu Arbeitsproduktivität, Reinertrag/AK und Rentabilitätskoeffizienten.

⇒ Abbildung B44 zeigt, daß diese Beziehung zwischen Lohn je entlohnte AK und Arbeitsproduktivität sehr deutlich ist. Fast 32% der Betriebe[45] mit durchschnittlicher oder überdurchschnittlicher Arbeitsproduktivität zahlen durchschnittliche oder überdurchschnittliche Löhne je AK, die verbleibenden gut 18% zahlen unterdurchschnittliche oder sehr geringe Löhne. Beim Begriff der unterdurchschnittlichen Arbeitsproduktivität tritt genau das entgegengesetzte Bild auf. Allerdings bleibt diese Betrachtung bei anderen Erfolgskennzahlen nicht erhalten. Abbildung B45 zeigt, daß beim Reinertrag/AK praktisch kein Zusammenhang zur Entlohnung der Arbeitskräfte mehr zu erkennen ist, daß zum Beispiel annähernd gleich viel Betriebe mit

---

[44] Diese Spitzengruppe resultiert aus einer Clusteranalyse der standardisierten Kennzahlen Betriebseinkommen/AK, Betriebeinkommen/Eqm und Betriebseinkommen in % BE, stellt also eine Gruppierung nach einem Kennzahlenmix dar. Als Clusteralgorithmus wird Ward verwendet.

[45] Richtiger ist es hier von Nennungen oder Fällen zu sprechen, da ja alle Jahre in die Prozentwerte eingehen, das heißt jeder Betrieb dreimal in den Werten auftaucht.

einem sehr hohen Reinertrag/AK unterdurchschnittlich (11,9% von 25% mit sehr hohem Reinertrag/AK) und durchschnittlich und überdurchschnittlich (13,1% von 25% mit sehr hohem Reinertrag/AK) entlohnen. Bei Betrachtung des Rentabilitätskoeffizienten ergibt sich sogar tendenziell ein leicht gegenteiliges Bild (Abbildung B46). Bei den Betrieben mit sehr hohen Rentabilitätskoeffizienten und Koeffizienten von über 1 überwiegen leicht die Betriebe, die unterdurchschnittlich entlohnt haben. Zoomt man in den Begriff mit sehr hohen Rentabilitätskoeffizienten und betrachtet nun die Beziehung Betriebseinkommen/AK zu Lohn/AK (Abbildung B47), so sieht man, daß 92,5% der Betriebe mit sehr hohem Rentabilitätskoeffizienten ein durchschnittliches oder überdurchschnittliches Betriebseinkommen/AK besitzen und daß eine Beziehung zwischen Betriebseinkommen/AK und Lohn/AK auf dieser Stufe nicht mehr vorliegt. Als Schlußfolgerung mag demnach festgehalten werden, daß die oben gemachte Feststellung auf einer ganz bestimmten Definition dessen beruht, was als erfolgreich verstanden wird, und in Beziehung zur Arbeitsproduktivität wohl zutrifft, bezogen auf Reinertrag/AK oder Rentabilitätskoeffizient aber in den Daten der Jahre 1992 bis 1994 nicht festgemacht werden kann.

- Feststellung 2: „Mehr Glasfläche je Arbeitskraft führt nach den vorliegenden Ergebnissen nicht zu höherem Erfolg. Die höhere Entlohnung bei den Spitzenbetrieben steht also vermutlich nicht mit der Arbeitsintensität (Kehrwert von ‘Fläche zu Arbeit’-Verhältnis, Anmerkung des Verfassers) in Zusammenhang“ (BITSCH, 1994, Seite 178). Diese Aussagen beruhen im wesentlichen auf den Schlußfolgerungen aus den Kennzahlenwerten der bereits oben angesprochenen Erfolgsgruppen. Ohne erneut das Problem des Begriffes Erfolg anzusprechen, werden in den folgenden Liniendiagrammen in den Abbildungen B48 bis B50 die Beziehungen der Arbeitsintensität zu Betriebseinkommen/AK, Rentabilitätskoeffizient und Lohn je entlohnte AK dargestellt.

⇒ Die Abbildungen B48 und B49 zeigen, daß es - abweichend von der formulierten Feststellung 2 - eine recht deutliche Beziehung von Arbeitsintensität und Erfolg gibt, wenn als Erfolgsmaßstäbe die Arbeitsproduktivität und der Rentabilitätskoeffizient verwendet werden. So haben fast die Hälfte der Betriebe mit einem sehr weitem ‘Fläche zu Arbeit’-Verhältnis auch eine sehr hohe Arbeitsproduktivität (12,1% von 24,8%) und über die Hälfte der Betriebe mit sehr weitem ‘Fläche zu Arbeit’-Verhältnis einen Rentabilitätskoeffizienten von größer 1 (14,8% von 24,8%). Demgegenüber weisen nur gut ein Drittel der Betriebe mit einem unterdurchschnittlichen ‘Fläche zu Arbeit’-Verhältnis eine durchschnittliche oder überdurchschnittliche Arbeitsproduktivität oder einen Rentabilitätskoeffizienten von größer 1 auf. Die entgegengesetzten Beziehungen von Arbeitsproduktivität und Flächenproduktivität zu dem ‘Fläche zu Arbeit’-Verhältnis (siehe 3.2.2.4) können möglicherweise erklären, daß der Arbeitsintensitätseffekt in der Analyse von BITSCH,

1994, von untergeordneter Bedeutung ist[46]. Abbildung B50 schließlich läßt vermuten, daß tendenziell die Entlohnung zunimmt, wenn die Arbeitsintensität abnimmt, das heißt das ‘Fläche zu Arbeit’-Verhältnis weiter wird, obwohl die Beziehung nicht sehr ausgeprägt ist. Auch diese Beobachtung liefert eher eine Einschränkung der oben formulierten Feststellung.

– Feststellung 3: „Die Entwicklung des Anteils der Eigenproduktion am Betriebsertrag zeigt ... eine Zunahme der Spezialisierung mit dem Erfolg (...). Die Ausweitung von Handel oder Dienstleistung erweist sich für Betriebe mit indirektem Absatz folglich nicht als erfolgsfördernd“ (BITSCH, 1994, Seite 199). Es stellt sich also die Frage, ob es einen sichtbaren Zusammenhang zwischen dem Anteil der Eigenproduktion am Betriebsertrag (epertp) und Erfolgskennzahlen gibt. Die Liniendiagramme in den Abbildungen B51 (für überwiegend indirekt absetzende Betriebe) und B52 (für überwiegend direkt absetzende Betriebe) gehen dieser Frage nach.

⇒ Abbildung B51 zeigt zunächst, daß sich bei den überwiegend indirekt absetzenden Betrieben der Anteil der Betriebe mit eher hohen und eher niedrigen Eigenproduktionsanteilen in etwa die Waage halten (54,8% zu 45,2%). Allerdings haben die Betriebe mit einem durchschschnittlichen und überdurchschnittlichen Eigenproduktionsanteil immerhin in über der Hälfte der Fälle Rentabilitätskoeffizienten von größer 1 (28,2% von 54,8%), während dieser Anteil bei den Betrieben mit unterdurchschnittlichen Eigenproduktionsanteilen nur bei gut einem Drittel liegt (16,7% von 45,2%). Bei sehr hohem Eigenproduktionsanteil sind es sogar 16,8% von 28%, daß sind 60% der Betriebe, die einen Rentabilitätskoeffizienten von größer 1 haben. Für die überwiegend indirekt absetzenden Betriebe wird die oben gemachte Feststellung also bestätigt. Bei den überwiegend direkt absetzenden Betrieben, fällt zunächst auf, daß die große Mehrzahl der Betriebe unterdurchschnittliche Eigenproduktionsanteile aufweist (82,3% gegenüber 17,7%), und diese Gruppe nur in knapp einem Drittel der Fälle (27,1% von 82,3%) einen Rentabilitätskoeffizienten von größer 1 besitzt. Zwar überwiegen auch bei den überwiegend direkt absetzenden Betrieben mit durchschnittlichem und überdurchschnittlichem Eigenproduktionsanteil die Betriebe mit Rentabilitätskoeffizienten von unter 1, der Anteil der erfolgreicheren Betriebe ist auf dieser Seite mit 7,7% von 17,7% aber doch spürbar höher, so daß die untersuchte Feststellung auch für die überwiegend direkt absetzenden Betriebe getroffen werden kann. Allerdings ist die Bildung der Variablenklassen durch die sehr viel größere Gruppe der überwiegend indirekt absetzenden Betriebe bestimmt und von daher für die überwiegend direkt absetzenden Betriebe etwas irreführend, da sie sich auf einem allgemein geringeren Niveau des Eigenproduktionsanteils bewegen. Schließlich fällt beim Vergleich der beiden Liniendiagramme noch auf, daß der Anteil der erfolgreicheren Betriebe (mit

[46] Obwohl hier noch einmal darauf hingewiesen werden muß, daß es sich in dieser Arbeit um andere Erhebungsjahre als in der Arbeit von BITSCH, 1994, handelt.

Rentabilitätskoeffizienten größer 1) bei den überwiegend indirekt absetzenden Betrieben mit 44,9% um immerhin 10% über den 34,8% der eher erfolgreichen direkt absetzenden Betriebe liegt.

- Feststellung 4: „ ... ein Zusammenhang zwischen der Betriebsgröße und dem Betriebserfolg kann nicht ausgeschlossen werden. *Als Hypothese läßt sich formulieren, daß Spitzenbetriebe größer sind als andere*“ (BITSCH, 1994, Seite 164, Kursiv durch Orginalautorin). Die Beobachtung, daß es offensichtlich Zusammenhänge zwischen Betriebsgröße und Erfolgskennzahlen gibt, wird ja bereits an anderen Stellen in dieser Arbeit angesprochen.

⇒ Die Liniendiagramme in den Abbildungen B53 und B54 zeigen, daß diese Feststellung im wesentlichen sowohl für die Arbeitsproduktivität gilt, wenn der Bezugsmaßstab die Glasfläche ist, als auch für die Flächenproduktivität gilt, wenn als Bezugsmaßstab für Größe die Anzahl der Arbeitskräfte gewählt wird. Werden die Bezugsmaßstäbe für den Begriff der Betriebsgröße vertauscht zeigen sich bei der Arbeitsproduktivität kaum noch (Abbildung B55) und bei der Flächenproduktivität eher entgegengesetzte Beziehungen (Abbildung B56). Wird schließlich der Rentabilitätskoeffizient betrachtet (Abbildungen B57 und B58), halten sich bei durchschnittlich und überdurchschnittlich großen Betrieben die Anteile der Betriebe mit Rentabilitätskoeffizienten von größer beziehungsweise kleiner 1 die Waage (bei beiden Bezugsmaßstäben), der Anteil der weniger erfolgreichen Betriebe ist bei den unterdurchschnittlich großen Betrieben aber deutlich größer als der Anteil der erfolgreicheren Betriebe. Von den 24,6% der Betriebe mit sehr niedrigen Rentabilitätskoeffizienten besitzen über 2/3 unterdurchschnittlich große Glasflächen und unterdurchschnittlich viele Arbeitskräfte (16,7% beziehungsweise 16,6% von 24,6%); von den 24% der Betriebe mit sehr hohen Rentabilitätskoeffizienten dagegen liegen 14,8% beziehungsweise 13% auf der Seite der größeren Glasflächen beziehungsweise der höheren Anzahl an AK. Die Liniendiagramme mit dem Rentabilitätskoeffizienten bestätigen also im wesentlichen die oben gemachte Feststellung, obwohl es sinnvoll erscheint davon zu sprechen, daß kleinere Betriebe eher zu den weniger erfolgreichen als zu den mehr erfolgreichen Betrieben zählen, während bei größeren Betrieben der Anteil erfolgreicher und weniger erfolgreicher Betriebe annähernd gleich ist. Viele weitere Fest- oder Fragestellungen ließen sich noch formulieren und mit Hilfe der interaktiven Liniendiagramme erkunden. Der grundsätzliche Weg ist aber durch die Beispiele ausreichend erläutert.

b-Identifikation von Gruppen und einzelnen Betrieben und deren Beschreibung

Das Programm TOSCANA, das die interaktive Exploration von Liniendiagrammen ermöglicht, eröffnet die Möglichkeit einzelne Objekte oder Gruppen von Objekten zu identifizieren und gesondert zu betrachten. Zoomt man in einen ausgewählten Begriff hinein, so wird das Liniendiagramm der darunterliegenden Ebene sichtbar. Der Umfang des Unterbegriffes entspricht dem des ausgewählten Oberbegriffes, und kann nach weiteren Inhalten untersucht werden kann. Dieser Vorgang ist beliebig oft und mit beliebigen Liniendiagrammen wiederholbar, bis man zu dem

Begriff gelangt, den man speziell betrachten möchte. Jedem Begriff ist die Anzahl der zum Begriff zählenden Objekte, das heißt sein Umfang, zugeordnet. Über DDE kann TOSCANA mit einer beliebigen Datenbank verbunden werden (im Beispiel mit MS-Access). In TOSCANA werden dann an jedem Begriff nicht nur die Anzahl der Fälle, sondern auch die Fallnummern aus der Datenbank angezeigt. Durch Anklicken gelangt man zu einem Access-Formularblatt, das alle in der Datenbank enthaltenen Informationen zum ausgewählten Objekt am ausgewählten Begriff enthält. Eine an einem Begriff gefundene Gruppe von Objekten, das heißt der Begriffsumfang, kann in Form der Fallnummern in die Zwischenablage gespeichert werden. Aus MS-Access heraus ist es dann möglich ein eigens für diesen Zweck gestaltetes Genstat-Menu zu starten, das auf Wunsch zu der ausgewählten Gruppe zusammenfassende Statistiken und Graphiken produziert. *Die Genstat Codes zur Erstellung dieser Menus sind im Anhang Teil III zu finden.* Abbildung B59 zeigt den Weg durch drei Liniendiagramme zu dem Begriff der sehr hoher Arbeitproduktivität, sehr hoher Wertschöpfungsquote und sehr hoher Flächenproduktivität entspricht. Es wird eine Gruppe von 35 Betrieben identifiziert. Abbildung B60 enthält die Menüpunkte in Genstat, die zeigen, welche Auswertungsmöglichkeiten zur Zeit bestehen[47], und Abbildung B61 einige der Graphiken und eine Liste einiger zusammenfassender univariater Statistiken, die unter anderem zeigen, daß die ausgewählte Gruppe an Betrieben einen deutlich über dem Median liegenden Rentabilitätskoeffizient besitzt, durchschnittlich groß ist, eine recht geringes ‘Fläche zu Arbeit’-Verhältnis besitzt und leicht überdurchschnittliche Löhne je AK zahlt. Abbildung B62 zeigt zwei ausgewählte Betriebe aus dieser Gruppe. Die Daten sind verändert, so daß die gezeigten Betriebe nicht tatsächlich am Kennzahlenvergleich beteiligten Betrieben entsprechen und hier nur als Beispiel dienen. Eine weitere Illustration liefert die Abbildung B63. Die gezeigten interordinalskalierten Diagramme sind für die Analyse im Zusammenhang mit CHAID konzipiert. Jede beliebige Zusammenlegung an Klassen ist erreichbar, und der Verlauf des Klassifikationsbaums läßt sich in den Liniendiagrammen nachvollziehen und verändern, Begriffe können ausgewählt und ihr Umfang im Detail betrachtet und wiederum über Genstat deskriptive Statistiken nachgefragt werden.

[47] Der Umfang der angebotenen Auswertungen ist beliebig erweiterbar.

4 Diskussion der Ergebnisse und Schlußfolgerungen

Der vorliegenden Arbeit liegt die Erwartung zugrunde, daß mit Hilfe explorativer datenanalytischer Verfahren einerseits aussagekräftige graphische Repräsentationen multivariater Daten geschaffen werden und andererseits inhaltliche Zusammenhänge aufgespürt, Hypothesen aufgestellt, voreilige Schlüsse vermieden und interessante Strukturen aufgedeckt werden können. Kapitel 4.1 faßt daher die inhaltlichen Ergebnisse aus Kapitel 3 noch einmal zusammen und formuliert Fragestellungen, die Anlaß zu weiteren Untersuchungen geben. Darüber hinaus ist eine Einschätzung der eingesetzten Methoden im Hinblick auf ihre 'Nützlichkeit' bei der Analyse der vorliegenden Daten erforderlich. Kapitel 4.2 faßt daher die methodischen Ergebnisse aus Kapitel 3 zusammen und diskutiert Stärken und Schwächen, sowie Möglichkeiten und Grenzen des eingesetzten Methodenspektrums, sofern dies nicht schon in Kapitel 2 geschehen ist. Kapitel 4.3 schließlich befaßt sich mit den Grenzen der vorgelegten Arbeit. Es erfolgt eine kritische Betrachtung des gewählten Vorgehens und eine Ausarbeitung von Vorschlägen für an diese Arbeit anschließende mögliche Fragestellungen und Forschungsprojekte.

4.1 Diskussion der inhaltlichen Ergebnisse

4.1.1 Betriebsbegleitende Untersuchung bei Cyclamen

Ziel in der Auswertung der Daten der betriebsbegleitenden Untersuchung bei Cyclamen (3.1) ist die Suche nach aussagekräftigen Darstellungsformen der gewonnenen Daten, die sowohl interessante Beobachtungen aufdecken als auch Diskussionen über die inhaltlichen Ergebnisse anregen, ohne sich von einzelnen auffälligen Werten blenden zu lassen. Alle gewählten Verfahren werden also in dem Sinne eingesetzt, daß sie zu einer Visualisierung der Daten beitragen. Das eine derartige Untersuchung nicht zu allgemeinen Schlußfolgerungen zur betrachteten Kultur dienen kann, wird bereits in Kapitel 3.1.1 erwähnt. Ein allgemeingültiger, erklärender Charakter der Beobachtungen wird nicht erwartet. Das die Auswahl der Verfahren notwendigerweise unvollständig bleiben muß, wird schon in der allgemeinen Einführung angesprochen. Die Methodenwahl folgt einer subjektiven Einschätzung des Verfassers, wie die Visualisierung und Analyse am besten vorangebracht werden kann. Als wichtige Ergebnisse der unterschiedlichen graphischen Verfahren treten hervor:

als allgemeine Ergebnisse in einer Vielzahl von Abbildungen:

- die Abbildungen lassen die Positionierung der beteiligten Betriebe zueinander und ansatzweise eine Gruppierung der Betriebe, bezogen auf unterschiedliche Merkmalsgruppen, erkennen;
- Betriebe mit auffälligen Werten in den verschiedenen Variablensets können leicht identifiziert werden;
- Variablenkorrelationen werden durch die Graphiken sichtbar und in ihrer Größenordnung

einschätzbar.

Auf dem Gebiet der Qualitätsbeurteilungen:

- zwischen Betrieben und Sorten deutet sich eine Wechselwirkung an (3.1.2.1 a-);
- im Durchschnitt erhält die Sorte ‘Sierra’ bessere Qualitätsbeurteilungen als die Sorte ‘Concerto’ (3.1.2.1 a-);
- in der Haltbarkeitsprüfung findet eine deutliche Verschlechterung der Qualitätsbeurteilung statt (3.1.2.1 a-);
- die Bonituren der einzelnen bonitierten Merkmale weisen innerhalb und zwischen den Betrieben erhebliche Schwakungen auf (3.1.2.1 a-);
- es gibt nur wenige Betriebe mit einer sehr guten oder sehr schlechten, und in der Regel nicht stabilen, Qualitätsposition (stabil über die beiden betrachteten Sorten beziehungsweise die beiden verwendeten Boniturzeitpunkte); viele Betriebe liegen dagegen im mittleren Qualitätsbereich (3.1.2.1 a-);
- die Korrelation zwischen der Beurteilung des Gesamteindrucks und der Beurteilung anderer Qualitätsmerkmale ist vergleichsweise gering (3.1.2.1 a-);
- die Abnahme der Polarisation des Mittels bei beiden Sorten während der Haltbarkeitsprüfung verdeutlicht die oben gemachte Aussage von der Abnahme der Qualitäten während der Haltbarkeitsprüfung (3.1.2.1 b-);
- die Beurteilung der Wurzelqualität weicht auffällig von den Beurteilungen der übrigen Qualitätsmerkmale ab, vor allem bei ‘Concerto’ (3.1.2.1 b- und 3.1.2.1 c-);
- die Veränderung der Korrelationsverhältnisse zwischen den beiden Qualitätsbeurteilungswochen läßt auf die Wirkung einer innerer Qualität der Pflanzen schließen, die in der Haltbarkeitsprüfung zur Geltung kommt (3.1.2.1 c-);
- die Ähnlichkeitsverhältnisse der Betriebe zueinander sowohl innerhalb der Sorten als auch innerhalb der Beurteilungswochen sind relativ stabil, allerdings bei ‘Concerto’ etwas stabiler als bei ‘Sierra’ (über beide Wochen) und in Woche 44 stabiler als in Woche 48 (über beide Sorten) (3.1.2.1 c-);
- dagegen sind die Ähnlichkeitsbeziehungen zwischen den Betrieben bei gleichzeitiger Betrachtung beider Sorten und beider Beurteilungswochen, daß heißt der Kombinationen ‘Sierra’ Woche 44, ‘Sierra’ Woche 48, ‘Concerto’ Woche 44, ‘Concerto’ Woche 48 gering, was den Schluß auf vier insgesamt recht unterschiedliche Konfigurationen zuläßt (3.1.2.1 c-).

Auf dem Gebiet der Substratanalysewerte:

- die unterschiedlichen Probenahmezeitpunkte weisen nur geringe Korrelation von Substratanalysewerten auf (3.1.2.2 a-);
- schwach korreliert sind auch die Substratanalysewerte der einzelnen untersuchten Nährstoffe, zumindest an zwei von drei Probenahmezeitpunkten (3.1.2.2 a-);
- auf eine Beziehung von hohen Substratanalysewerten und geringen Werten bei den Qualitätsbonituren geben die vorhandenen Werte nur einen schwachen Hinweis (3.1.2.2 a-);

Auf dem Gebiet der Kulturmaßnahmen:

- Hauptunterscheidungsmerkmale zwischen den Betrieben in Bezug auf ihre Steuerung der Schattierung sind die absolute Stärke der Schattierung und die Veränderung der Schattiersollwerte im Kulturverlauf (3.1.2.3 a-);
- Betriebe mit dunkler Kulturführung gehören vermehrt zu den Betrieben mit geringer Qualität (3.1.2.3 a-);
- eine Gruppierung der Betriebe mit Hinblick auf ihre Kulturmaßnahmen wird vor allem durch Unterschiede im Energieverbrauch, Platzbedarf und Rücken unterstützt (3.1.2.3 b-);
- eine sehr deutliche und klar abgegrenzte Gruppierung der Betriebe ist jedoch nicht möglich, da selbst Betriebe mit einer noch relativ großen Ähnlichkeit, vielfach dennoch erhebliche Unterschiede bei den Ausprägungen der einzelnen Merkmale besitzen (3.1.2.3 b-).

Auf dem Gebiet der Strukturmerkmale:

- flächenmäßig größere Betriebe produzieren weniger Cyclamen als flächenmäßig kleinere Betriebe (3.1.2.4 a- und 3.1.2.4 b-);
- Betriebe mit modernen Stellflächen und Bewässerungsverfahren von unten setzen überwiegend Einheitserden ein (3.1.2.4 a- und 3.1.2.4 b-).

Beim simultanen Vergleich aller Variablensets:

- Die Ähnlichkeitsbeziehungen in den einzelnen Variablensets weichen stark voneinander ab (3.1.2.5 a-);
- Erfolg oder Mißerfolg in der Kultur kann nicht durch die bestimmten Kulturmerkmale in ihrer Gesamtheit erklärt oder verstanden werden (3.1.2.5 a-);
- Endstand und mittlere Stärke der Schattierung zählen zu den am stärksten diskriminatorischen

Variablen zwischen den Betrieben (3.1.2.5 b-);

- Kulturmerkmale und Qualitätsbeurteilungen stehen nur in einem losen und nicht für beide Sorten und Beurteilungszeitpunkte einheitlichen Beziehungszusammenhang (3.1.2.5 b-);
- in Einzelfällen ergeben sich zu den Beziehungen der Variablensets untereinander inhaltlich interessante Aussagen, zum Beispiel für die Beziehungen bei ‘Concerto’ in Beurteilungswoche 44, ohne daß sich diese Erklärungsmuster in den anderen Varianten wiederholen (3.1.2.5 c-).

Die Auflistung zeigt, daß sich aus der Analyse der Daten eine Vielzahl Feststellungen ergibt, die in fast allen Fällen direkt aus der Betrachtung der gewählten Graphiken resultieren. Insofern wird das Ziel einer reichhaltigen und informativen Visualisierung erfüllt.

Einzelne Beobachtungen geben darüber hinaus Anregungen gezielte Versuche durchzuführen, zum Beispiel zur Steuerung der Schattierung, dem optimalen Endstand oder dem vorteilhaftesten Rückverhalten.

Allerdings verdeutlicht die betriebsbegleitende Untersuchung auch, daß einerseits die verschiedensten Kulturbedingungen in der Lage sind, vergleichbare Qualitäten zu produzieren, und andererseits sehr unterschiedliche Qualitäten aus Betrieben mit ähnlichen Kulturführungen stammen können. Effekte scheinen sich also gegeneinander aufzuheben, und partielle Rezepte, die in einem Betrieb zum Kulturerfolg führen, tun dies im anderen Betrieb noch lange nicht. Betriebsbegleitenden Untersuchungen kann in der Aufdeckung dieser Wechselwirkungen der Kulturbedingungen eine große Bedeutung zukommen, wenn es um die Verstehen von Kulturabläufen geht. Durch die Erfassung der Reaktion der Kultivateure auf unvorhergesehene Ereignisse läßt sich das Expertenwissen der Gärtner formalisieren. Eine erhebliche Grenze Neues durch betriebsbegleitende Untersuchungen zu lernen ist allerdings dadurch gegeben, daß jeder Betrieb, die ihm zur Verfügung gestellten Pflanzen so gut wie möglich kultivieren und letztendlich auch verkaufen will. Experimente in Betrieben, bei denen die Pflanzen Schaden nehmen könnten, ließen sich daher wohl nur mit zusätzlichen finanziellen Anreizen durchführen. Die Nachkontrolle experimenteller Ergebnisse in der betrieblichen Wirklichkeit durch betriebsbegleitende Untersuchungen kann aber zu einem erheblichen Zugewinn an Wissen über das Verhalten der Pflanzen und der Produzenten im Erzeugungsprozeß führen. Allerdings sind dann einige Anforderungen an die betriebsbegleitenden Untersuchungen zu stellen, die durch hier besprochene Untersuchung mit Cyclamen, nicht erfüllt werden. Dazu zählen: deutliche Erhöhung der Anzahl der beteiligten Betriebe; ständige Kontrolle und quantitative Erfassung von Klima- Ernährungs- und Wachstumsparametern; und Objektivierung der Qualitätsbeurteilung während und nach der Kultur.

Viele im Zusammenhang mit der betriebsbegleitenden Untersuchung gewonnenen Ergebnisse lassen zwei Schlußfolgerungen zu. Entweder sind Cyclamen relativ unempfindlich gegenüber unterschiedlichen Kulturmaßnahmen und -eingriffen und wachsen unter einer Vielzahl von

Bedingungen zu einem vergleichbaren Produkt heran, oder die Reaktion der Kultivateure sorgt für das entsprechende Zusammenspiel der Wachstumsfaktoren, so daß ein jeweils auf die betrieblichen Bedingungen abgestimmtes Umfeld entsteht, das je nach Reaktionsvermögen des Produzenten zu einer letztlich befriedigenden oder unbefriedigenden Qualität führt. Ein Beispiel hierfür könnten die Betriebe 12 und 14 liefern. Bei ‘Concerto’ am Kulturende erhalten beide Sorten mit einer Ausnahme übereinstimmende, relativ hohe, Boniturwerte (Abbildungen A3 und A9). In den einzelnen Variablensets weisen diese Betriebe stark voneinander abweichende Positionen auf, zum Beispiel bei der Ernährung (Abbildung A32), der durchschnittlichen Stärke der Schattierung (Abbildung A39 a)), der Gruppierung nach Kulturmaßnahmen (Abbildung A44 und A45), und den Strukturdaten Stellfläche und Bewässerungsverfahren (Abbildungen A52 und A54). Die Vermutung, die unterschiedliche Kulturführung schlage sich in einem veränderten Verhalten in der Haltbarkeitsprüfung nieder, mag durch die insgesamt etwas bessere Qualitätsbeurteilung von ‘Concerto’ bei Betrieb 14 in Woche 48 auf den ersten Blick eine Bestätigung finden. Bei ‘Sierra’ ist jedoch weder die Nähe der beiden Betriebe im Hinblick auf ihre Qualität zu beobachten (bei gleicher Kulturführung wie bei ‘Concerto’), noch eine vergleichbare Umpositionierung von Woche 44 auf Woche 48 zu beobachten.

Dies ist aber ein Einzelbeispiel, das nicht zur Verallgemeinerung Anlaß gibt. Um mehr durch das Vorgehen des Produzenten über Wachstum und Entwicklung von Cyclamen (und anderen Zierpflanzen) zu erfahren, ist eine Veränderung der betriebsbegleitenden Untersuchungen in der oben skizzierten Form ratsam. Erst dann kann geklärt werden, welche nachvollziehbaren Gesetzmäßigkeiten sich hinter der in dieser Arbeit beobachteten Variabilität verbergen, und ob der Kulturerfolg überhaupt, und wenn ja in welchen Toleranzbereichen und durch welche Kombination von Kulturbedingungen, durch die Kulturmaßnahmen beeinflußt wird. Da von einer derartigen Beeinflussung ausgegangen werden kann, die Klärung der Zusammenhänge aber sehr komplex ist, könnte ein gezielter und geplanter Ausbau betriebsbegleitender Untersuchungen mit exakten Aufzeichnungen sicherlich das Verständnis nicht nur für die Cyclamenkultur verbessern.

4.1.2 Kennzahlenvergleich

Während auch bei der Auswertung der Kennzahlen das Bemühen um aussagekräftige Formen der graphischen Darstellung in dieser Arbeit eine große Bedeutung haben, spielt die Besprechung und Analyse der inhaltlichen Zusammenhänge eine größere Rolle als in der Auswertung der betriebsbegleitenden Untersuchung, obwohl auch in diesem Kapitel darauf hingewiesen werden muß, daß es sich bei den Kennzahlen um eine wenig vollkommene Datengrundlage handelt, und viele der betrachteten Werte mit Aufzeichnungsungenauigkeiten und ähnlichen Defekten behaftet sein können[48]. Die Auswahl der Verfahren richtet sich daher im wesentlichen an konkreten Fragestellungen aus, die auch in den einzelnen Kapiteln genannt werden, wobei das Augenmerk in der Hauptsache auf Methoden gelegt wird, die in unterschiedlichen Formen Zusammenhänge zwischen Kennzahlen durch graphische Methoden transparent zu machen in der Lage sind. Es ist aber erneut zu betonen, daß es nicht das Ziel dieser Arbeit ist, eine komplette und in sich geschlossene Analyse der Kennzahlen zu liefern, sondern Anstöße zu geben, sich dem Kennzahlenkatalog auf unterschiedliche Weise zu nähern und exemplarisch darzustellen, wie durch die eingesetzten Verfahren Erkenntnisse gesammelt und Hypothesen formuliert werden können. Die inhaltlichen Feststellungen können daher sowohl bekannte Tatsachen bestätigen, als auch weniger beachtete Fragestellungen und Hypothesen aufwerfen, für die Erklärungsmuster bereits vorhanden sind oder auch nicht. Eine inhaltliche Auseinandersetzung mit den Kennzahlen ist also nicht Ziel dieser Arbeit und bleibt den Gartenbauökonomen vorbehalten. Es kann aber gezeigt werden, daß die beschriebenen explorativen datenanalytischen Verfahren in der Lage sind, Zusammenhänge nachvollziehbar zu untersuchen und darzustellen. Ohne erneut auf alle einzelnen Beobachtungen einzugehen, lassen sich einige der Ergebnisse aus den fünf Auswertungsschritten wie folgt zusammenfassen:

- die Klassifizierung der Betriebe nach Anzahl AK, Glasfläche und Unternehmensertrag führt zu ungleich besetzten Klassen. Eine Verwendung gleich großer Klassen, eventuell sogar mit Überlappungsbereichen, führt zu einer besseren Vergleichbarkeit der Gruppen. Es wird allerdings eingestanden, daß durch den Verlust an Kontinuität von Jahr zu Jahr die Vergleichbarkeit der Gruppen im Zeitablauf verschlechtert wird (3.2.2.1).
- Viele Kennzahlen besitzen Ausreißer und extreme Werte; arithmetischer Mittelwert und Median zeigen häufig große Unterschiede; in dieser Arbeit daher wird zur Beschreibung der zentralen Tendenz ausschließlich der Median und nicht das arithmetische Mittel verwendet (3.2.2.1).
- Die Korrelationen zwischen Kennzahlen des Erfolgs oder der Produktivität (die in den

[48] Schließlich entstammen die Kennzahlen den steuerlichen Bilanzen, die ja in erster Linie unter steuerlichen und gesetzlichen Gesichtspunkten erstellt werden.

Auswertungen als Erfolgskennzahlen zusammengefaßt werden) und anderen Kennzahlen sind bei der Mehrzahl der ausgewählten Kennzahlen gering. Es stellt sich demnach die Frage, ob es nicht stärker mit den Erfolgskennzahlen in Zusammenhang stehende betriebliche Kennwerte gibt, deren Ermittlung sinnvoll wäre. Diese sind aber möglicherweise nicht den Buchführungsabschlüssen zu entnehmen (zum Beispiel Preisniveau des Absatzweges, Persönlichkeitsprofil des Unternehmers) (3.2.2.1).

- Obwohl die absoluten Werte in nach den Gesichtspunkten Betriebsgröße, Region und Erhebungsjahr gruppierten Kennzahlen durchaus voneinander abweichen, gibt es zwischen den so gruppierten Daten eine große Übereinstimmung hinsichtlich ihrer wesentlichsten Varianzursachen und der am stärksten diskriminatorischen Variablen. So sind es in erster Linie die Erfolgskennzahlen und prozentualen Aufwandswerte, und an zweiter Stelle Vermögens- und im weitesten Sinne technologiebezogene Kennzahlen, die zur Unterscheidung der Betriebe beitragen (3.2.2.2).

- Erhebliche Unterschiede bei den absoluten Werten in den Gruppierungen nach Betriebsgröße, Region und Erhebungsjahr treten vor allem bei den unterschiedlichen Regions- und Betriebsgrößenklassen auf, während die Ergebnisse zwischen den Erhebungsjahren große Überschneidungen aufweisen. Eine Abnahme der Regionsunterschiede mit Zunahme der Betriebsgröße deutet sich an (3.2.2.2).

- Eine Segmentierung der am Kennzahlenvergleich beteiligten Betriebe auf Grundlage der Beziehung einzelner Kennzahlen zu einem Erfolgskriterium, dem Rentabilitätskoeffizienten, identifiziert die prozentualen Aufwandswerte als die am stärksten segmentierenden Kennzahlen. Nur bei Ausschluß dieser Werte treten Betriebsgröße, Fremdkapitalanteil und regionale Lage als weitere stark segmentierende Variablen hervor (3.2.2.3).

- Bezogen auf den Rentabilitätskoeffizienten stellt sich die Frage, wie die höhere Aufwandseffizienz der erfolgreicheren Betriebe erklärt werden kann. Die geringen Korrelationen der Aufwands- Vermögens- und Strukturkennzahlen untereinander deuten darauf hin, daß die in dieser Arbeit verwendeten Kennzahlen eine zufriedenstellende Beantwortung dieser Frage nicht liefern können (3.2.2.3).

- Es stellt sich die Frage, ob diese Tatsache überhaupt mit den vorhandenen Kennzahlen erklärt werden kann, da höhere Aufwandseffizienz seine Ursache in besseren Preisen, geringeren Verlusten (in der Kultur und in der Vermarktung) oder in größeren Produktionsmengen haben kann. Zu diesen drei Punkten enthält der Kennzahlenkatalog keine Informationen. Insbesondere eine allgemeine Einstufung des Preisniveaus, unter dem der jeweilige Betrieb operiert, könnte in diesem Zusammenhang einen erheblichen Informationsgewinn darstellen (3.2.2.3).

- Unterschiedliche Erfolgskennzahlen spiegeln unterschiedliche Beziehungen zu anderen Kennzahlen wieder. Insofern sollte eine Eingruppierung der Betriebe auf Grundlage einzelner

Kennzahlen in Gruppen erfolgreicher und weniger erfolgreicher Betriebe immer berücksichtigen, daß andere Kriterien zu anderen Eingruppierungen führen können und damit auch andere Zusammenhänge sichtbar machen würden (3.2.2.4).

- Auffällig in diesem Zusammenhang sind zum Beispiel die gegenläufigen Beziehungen von Arbeits- und Flächenproduktivität zu Kennzahlen der Betriebsgröße (Eqm). Flächenmäßig größere Betriebe zählen eher zu den Betrieben mit hoher Arbeitsproduktivität, flächenmäßig kleinere Betriebe zu Betrieben mit höherer Flächenproduktivität. Der Zuwachs der Flächenproduktivität hält demnach nicht mit dem Zuwachs an Fläche mit. Es ist zu fragen, woran dies liegt (3.2.2.4).

- Die deutlich positive Beziehung zwischen Lohn je entlohnte AK und Betriebseinkommen/AK setzt sich nicht auf andere Kennzahlen, wie zum Beispiel den Rentabilitätskoeffizienten oder den Reinertrag/AK fort. Bevor also Aussagen zu den Beziehungen von Kennzahlen zueinander verallgemeinert werden, bietet sich eine vielschichtige Überprüfung des scheinbar ermittelten Zusammenhangs an (3.2.2.5).

- Beim Bemühen um eine Diskretisierung der Kennzahlen fällt auf, daß nur die wenigsten der betrachteten Kennzahlen, eine absolute eigene inhaltliche Bedeutung haben, daß heißt, was viel oder wenig, groß oder klein, erfolgreich oder nicht erfolgreich, produktiv oder unproduktiv ist, definiert sich ausschließlich aus dem Vergleich mit andern Betrieben. Wenn also der Kennzahlenkatalog erneuert oder umgeschrieben werden sollte, so ist darauf zu achten, daß Kennzahlen entwickelt werden, die als solche eine direkte inhaltliche Bedeutung besitzen, wie zum Beispiel der Rentabilitätskoeffizient, der beim Schwellenwert von 1 eine klare Interpretation liefert (3.2.2.5).

- Auf der anderen Seite ist es eventuell sogar möglich aus den vorhandenen Kennzahlen derartige inhaltlich relevante Schwellenwerte zu extrahieren. Zwei Beispiele, die in dieser Arbeit in diese Richtung weisen, sind die Beziehung der prozentualen Aufwandskennzahlen, vor allem der Lohnquote, zum Rentabilitätskoeffizienten bei einer Segmentierung der Kennzahlenbetriebe und das Beziehungsgeflecht von Betriebsgröße, Anzahl Arbeitskräfte und ‘Fläche zu Arbeit’-Verhältnis (3.2.2.3 und 3.2.2.4).

Die genannten Punkte sind als Diskussionsanregungen gedacht. Im Rahmen dieser Arbeit ist nicht beabsichtigt, die inhaltliche Richtigkeit und Relevanz aller Einzelheiten abschließend zu beurteilen. Es wird aber an verschiedenen Stellen der Auswertung der Kennzahlen verdeutlicht, daß unterschiedliche Betrachtungsweisen derselben Daten zu unterschiedlichen Schlußfolgerungen führen, und daß daher von einer Verallgemeinerung einzelner Auffälligkeiten Abstand genommen werden soll. Auch wird der eingeschränkte Informationsgehalt der Kennzahlen zur Klärung allgemeingültiger betriebswirtschaftlicher Zusammenhänge erneut unterstrichen (siehe zum

Beispiel auch BITSCH, 1994). Forderungen für eine Weiterentwicklung des Kennzahlenvergleichs, die sich aus dieser Arbeit ergeben, sind:

1. Der Median sollte statt des Mittelwerts zur Beschreibung der zentralen Tendenz verwendet werden;
2. die vorhandenen Kennzahlen sollten durch Kennwerte ergänzt werden, die das betriebliche Umfeld charakterisieren;
3. wenn eine Gruppierung der Betriebe vorgenommen werden soll, sollte über die Einführung flexibler Gruppierungsverfahren wie CART oder CHAID zur Identifikation homogener Gruppen an Stelle der traditionellen Gliederung des Arbeitskreises Betriebswirtschaft nach Betriebseinkommen/AK nachgedacht werden ;
4. durch die Bereitstellung von interaktiven Werkzeugen, wie zum Beispiel von hierarchischen Liniendiagramme zur Erkundung der Datengrundlage, oder Beantwortung konkreter Fragestellungen, könnte die Transparenz der Daten auch für die die Daten nutzenden Berater verbessert werden;
5. quantitative Kennzahlen sollten, wo sinnvoll, durch qualitative Kennzahlen, die auch aus einer Klassenbildung bei quantitativen Kennzahlen hervorgehen kann, ersetzt werden (dieser Punkt wird im folgenden Kapitel noch einmal angesprochen);
6. dazu wäre allerdings die Entwicklung neuer Kennzahlen erforderlich, die diese sinnvolle Klassenbildung überhaupt erst ermöglichen. Darüberhinaus sollten aussagekräftiger Schwellenwerte bei schon vorhandenen Kennzahlen entwickelt werden.

4.2 Diskussion der Methoden

Wo die Möglichkeiten und Grenzen der eingesetzten Methoden zur Datenanalyse liegen wird, ausführlich in Kapitel 2 dargestellt. Die dort erarbeiteten grundsätzlichen Aussagen bilden die Grundlage aller Auswertungen in Kapitel 3. In der Auswertung kann die Vielfältigkeit der Methoden demonstriert werden. Eine Beurteilung im Sinne einer guten oder einer schlechten Methode ist allerdings nicht möglich. Nur im Einzelfall kann entschieden werden, was gezeigt oder untersucht werden soll, und ob ein Verfahren zur gewünschten Darstellung geeignet ist oder nicht. In der vorgelegten Arbeit lassen sich drei Hauptgruppen von datenanalytische Ansätzen, die einen breiten Raum in der Auswertung und Darstellung der Daten einnehmen, unterscheiden. Erstens Visualisierung, zweitens Gruppierung und Segmentierung und drittens Untersuchung von Beziehungsgefügen von Variablen[49].

4.2.1 Verfahren zur Visualisierung - Biplots

Die Verfahren der Dimensionserniedriegung und ihre Darstellung in Form von Biplots (Kapitel 2.1 und 2.2) dienen der Sichtbarmachung von Informationen. Welches Verfahren gewählt wird, hängt von der vorhandenen Datenstruktur und der erwünschten Aussage ab. Bis zu einer bestimmten Anzahl an Merkmalen und Objekten ist die Biplotdarstellung in den Beispielen sinnvoll möglich. In der Auswertung der betriebsbegleitenden Untersuchung stößt man mit der Biplotmethodik nie an eine ernstzunehmende Darstellungsgrenze. Schwieriger wird es schon in der Analyse der Kennzahlen. Die Vielzahl an Objekten und Variablen, macht eine gemeinsame Darstellung schwierig. Eine Aufteilung in Gruppen oder eine Aufsplittung in Variablen- und Objektplots ist dann vorteilhaft, obwohl, vor allem für den Vergleich von Gruppen, eine visuelle Inspektion der Biplots alleine nicht mehr ausreicht um Gruppenunterschiede zu überprüfen, sondern formale Verfahren zur Untersuchung unterschiedlicher Gruppen in Anspruch genommen werden. Der Informationsgehalt aller Biplots ist hoch, wenn auch einschränkend festgestellt werden muß, daß mit der zweidimensionalen Approximation immer ein mehr oder weniger großer Informationsverlust einhergeht. Um die Güte der Abbildung einzuschätzen werden Screeplot, CUSUM-Diagramm, überlagerte Multiple Spanning Trees, Residuenanalysen, Stabilitätsprüfungen und unterschiedliche Verfahren zur Ermittlung der Anzahl der ‘wesentlichen’ Komponenten vorgeschlagen. Die in vielen Fällen erforderliche Standardisierung der Daten vor der Durchführung der Hauptkomponentenanlyse vermindert die Aussagekraft eines Hauptkomponentenbiplots, da es ja gerade die Ablesbarkeit der Orginalwerte an den Biplotachsen ist, die diese Darstellungsform so interessant macht. Mit Hilfe des entwickelten Genstat Codes wird dieses Problem allerdings überwunden. Unbefriedigend bleiben auch in dieser Arbeit die Darstellungsversuche in mehr als

[49] Natürlich gibt es zwischen diesen Ansätzen erhebliche Überschneidungen. So liefert ein Biplot auch Informationen über Variablenbeziehungen, fällt aber in der hier bezeichneten Gliederung unter den Oberbegriff Visualisierung, da Biplots in dieser Arbeit hauptsächlich als Visualisierungsinstrument eingesetzt werden.

zwei Dimensionen, obwohl durch die Andrews Kurven und konditionierten Hauptkomponentenwerteplots recht ansprechende Abbildungen geschaffen werden können. *Insofern können die Methoden der Dimensionserniedrigung, so wie sie hier bearbeitet werden, als eine Bereicherung in der Darstellung von mehrdimensionalen Datensätzen in der gartenbaulichen Beratung angesehen werden, sofern, für den Fall, daß eine Identifizierung einzelner Objekte gewünscht wird, die Anzahl der Objekte nicht zu groß ist*

4.2.2 Verfahren zur Visualisierung - hierarchische Liniendiagramme

Hierarchische Liniendiagramme ermöglichen im wahrsten Sinne des Wortes eine Exploration umfangreicher Datensätze. Die Interaktivität der Liniendiagramme ermöglicht dem Benutzer beliebige Abfragekonstellationen herzustellen und die Daten nach Auffälligkeiten zu durchsuchen, ohne daß der direkte Bezug zu den Ausgangsdaten verloren geht. Obwohl theoretisch durch Liniendiagramme jeglicher Begriffsverband komplett und ohne Informationsverlust dargestellt werden kann, sind dem natürlich auch praktische Grenzen gesetzt. Desweiteren ist ihr Einsatz nur nach einer begrifflichen Skalierung möglich. Diese begriffliche Skalierung (oder auch die Klassenbildung für den Einsatz diskreter graphischer Modelle oder von CHAID) ist aber im Kontext der Kennzahlen durchaus sinnvoll und vertretbar. Es ist allerdings nicht so, daß die in dieser Arbeit gewählte begriffliche Skalierung die einzig richtige oder die wirklich beste ist. Sie soll aber den Anstoß dazu liefern eine begriffliche Skalierung für die Kennzahlen zu entwickeln. Nun mag eine Klassenbildung einerseits sehr subjektiv erscheinen und andererseits, die in den Kennzahlen enthaltenen Informationen verkürzen und zudem durch die Ziehung der Klassengrenzen eine bestimmte Willkürlichkeit einführen. Die Liniendiagramme lassen aber, wie an den Beispielen gezeigt wird, eine sehr detaillierte und vor allem auch inhaltlich begründete Klassenbildung, sowie auch die Betrachtung aggregierter Klassen und beliebiger Klassenkombinationen, zu, so daß bei einer konkreten inhaltlichen Definition der gewünschten Begriffe, die noch weit spezifischer sein können als in den Beispielen, entsprechende Liniendiagramme aufgebaut werden könnten. Die Klassenbildung entspricht darüberhinaus einer entsprechenden Transformation der Daten, die aufgrund der hohen Anzahl an extremen Werten beziehungsweise den erheblichen Abweichungen von der Normalverteilung für eine multivariate Betrachtung auch mit anderen Methoden (zum Beispiel graphischen Modellen, siehe unten) erforderlich ist. Schließlich spricht noch ein weiterer Punkt für die Diskretisierung der Kennzahlen. Die Interpretation der Kennzahlen verläuft in der Regel diskret, das heißt, wenn Kennzahlenergebnisse in der Literatur diskutiert werden, so sind es häufig nicht die einzelnen absoluten Werte, die hervorgehoben werden, sondern gewisse Bereiche, die als zufriedenstellend empfunden werden. Ihre Beurteilung wird in den meisten Fällen aus dem Vergleich mit (wie auch immer gewonnenen) Gruppenmittelwerten abgeleitet (BAHNMÜLLER, 1997 & 1998). Eine begriffliche Skalierung würde an dieser Stelle eingreifen und die verwendeten Beurteilungskriterien objektivieren können (wobei die Schaffung neuer Kennzahlen oder die Ermittlung entsprechender Schwellenwerte notwendig werden würde, siehe Abschnitt 4.1.2). *Da auch für die einzelbetriebliche Betrachtung hierarchische Liniendiagramme ein wertvolles Hilfsmittel sind, wenn es um die Durchforstung der Datengrundlage geht und sich gleichzeitig Informationen für die Gesamtheit der Kennzahlenbetriebe abgerufen werden können, lautet eine Anregung dieser Arbeit, ein begriffliches Schema in Zusammenarbeit von Gartenbauökonomen, dem Arbeitskreis Betriebswirtschft und der gartenbau-betriebswirtschaftlichen Beratung für den Kennzahlenvergleich aufzubauen.*

4.2.3 Verfahren zur Visualisierung - Trellis-Displays

Trellis-Displays schließlich bieten sich zur Darstellung und Erforschung einer Vielzahl möglicher Fragestellungen an. Es handelt sich bei ihnen um eine flexible und übersichtliche Ergänzung und Verfeinerung des klassischen Spektrums univariater Graphiken. Die Konditionierung durch qualitative Variablen (oder diskretisierte quantitative Variablen) läßt eine Vielzahl an Kombinationsmöglichkeiten und Datenzusammenfassungen zu. Der theoretisch beliebig fein strukturierten Konditionierung sind jedoch praktische Grenzen gesetzt, das heißt, man stößt natürlich auch mit Trellis-Displays bei Betrachtung einer zu großen Variablenzahl an einen Punkt, wo die Vielzahl an Informationen nicht mehr in einer einzelnen Abbildung vermittelt werden kann. Eine gleichzeitige Darstellung mehrer Trellis-Displays auf einer Seite, wie sie an verschiedenen Stellen dieser Arbeit verwendet wird, zeigt jedoch, daß die Darstellungsmöglichkeiten erheblich sind. Da in dieser Arbeit immer wieder deutlich wird, daß die Konditionierung durch diskretisierte, kontinuierliche Variablen (zum Beispiel in der Analyse der Kennzahlen Einheitsquadratmeter oder Anzahl AK) zur Aufdeckung unterschiedlicher Merkmalsbeziehungen in den einzelnen gebildeten Klassen führt, wird die Brauchbarkeit dieser Vorgehensweise unterstrichen. *Eine umfangreiche Einbeziehung von TrellisDisplays sowohl in der Erforschung von Zusammenhängen als auch in der Ergebnisdarstellung bietet sich daher sowohl in der Analyse betriebsbegleitender Untersuchungen als auch der Analyse der Kennzahlen an.*

4.2.4 Gruppierung und Segmentierung - Clusteranalyse

Clusterverfahren werden nur an zwei Stellen in dieser Arbeit eingesetzt. Dies in erster Linie, weil weder in der betriebsbegleitenden Untersuchung noch in der Auswertung der Kennzahlen nach homogenen Gruppen gesucht wird. Die Clusteranalyse dient aber gerade in erster Linie dazu, homogene Gruppen zu identifizieren. Die Ableitung kausaler Zusammenhänge aus einer Clusteranalyse, wie sie zum Beispiel von BITSCH, 1994, vorgenommen wird, ist demgegenüber mit großen Schwierigkeiten behaftet, da die Gruppenbildung in der objektorientierten Clusteranalyse nicht auf Variablenbeziehungen, sondern einem gewählten Homogenitätskriterium, in den hierarchischen Clusteranalysen zum Beispiel auf einem speziellen Proximitätsmaß, beruhen, und unterschiedlichste Merkmalsausprägungen zu identischen Proximitäten zwischen Objekten führen können. Wenn nach Variablenzusammenhängen auf Grundlage der gebildeten Cluster bei Variablen gesucht wird, die an der Bildung der Cluster überhaupt nicht beteiligt sind, wie dies bei BITSCH, 1994, erfolgt, scheint die Clusteranalyse nicht angemessen zu sein. BACHER, 1994, unterscheidet in diesem Zusammenhang zwischen Strukturgleichungsmodellen und Clusteranalyse und führt aus: „Aufgabe von Strukturgleichungsmodellen ist die Spezifizierung und/oder Überprüfung von kausalen Beziehungen zwischen Variablen, primäres Ziel von Clusteranalyseverfahren dagegen das Auffinden einer empirischen Klassifikation und unter Umständen das Auffinden einer hierarchischen Ähnlichkeitsstruktur.“ (BACHER, 1994, Seite 10). Darüber hinaus ist es auch bemerkenswert, daß trotz der vielen Weiterentwicklungen auf dem Gebiet der Clusteranalyse, die zwischen EVERITT, 1979, und ARABIE & HUBERT, 1995, liegen, viele Problembereiche nach wie vor nicht zufriedenstellend gelöst sind (gibt es überhaupt eine Clusterstruktur? wieviele Cluster liegen vor? welcher Clusteralgorithmus und welches Proximitätsmaß?). *Schließlich ist anzumerken, daß die in dieser Arbeit durchgeführten, sicher sehr unvollständigen Ansätze zur Clusteranalyse keine überzeugenden Hinweise dafür in den Daten gefunden haben, daß die betrachteten Objekte überhaupt eine Clusterstruktur besitzen.*

4.2.5 Gruppierung und Segmentierung - CART und CHAID

Klassifikations- und Regressionsbäume haben demgegenüber für die Aufgabenstellung dieser Arbeit eine größere Bedeutung, da durch ihre Konstruktion gleichzeitig mit der Bildung der Segmente auch eine Beschreibung der wichtigsten, die Segmente beschreibenden Variablen, entsteht und darüberhinaus eine sich selbst erklärende Darstellung der Zusammenhänge gebildet wird. Insofern erfüllen sie eher das Bedürfnis nach einer umfassenden Visualisierung der Bildung der Segmente als die Dendrogramme der Clusteranalyse. Auch liefern sie gerade bei derartig wenig perfekten Daten wie sie hier vorliegen, eine willkommene, robuste Alternative zur linearen Regressionsanalyse, wenn es um die Betrachtung von Beziehungen mehrerer erklärender und einer abhängigen Variablen geht. Vor allem die dadurch erzielte Transparenz bezüglich der Bildung der Segmente wird als wesentlicher Vorteil gegenüber der Clusteranalyse empfunden. Die unbeantworteten Fragen, die aber auch in diesem Methodenbereich noch liegen, dürfen allerdings nicht übersehen werden, so zum Beispiel die Frage nach der besten Splitting-Regel, der optimalen Baumgröße, den Entscheidungungskriterien beim pruning oder auch die Verwendung von Teststatistiken (siehe zum Beispiel LOH & VANICHSETAKUL, 1988, PANEL, 1989, NAGEL et al., 1996). Diese Problembereiche werden in dieser Arbeit jedoch nicht thematisiert, sondern es wird ein ausgesprochen pragmatischer und beispielsorientierter Weg beschritten. Dabei wird deutlich, daß die Klassifikations- und Regressionsbäume zu gut interpretierbaren Abbildungen führen, die einerseits durch ihre Einfachheit und Nachvollziehbarkeit beeindrucken, andererseits aber auch dazu führen können, den Eindruck einer Eindeutigkeit zu vermitteln, die bei weitem durch die Daten nicht gedeckt ist. *Dennoch kann eine auf einer CART oder CHAID Methodik begründete Segmentierung der Kennzahlenbetriebe unter Verwendung unterschiedlicher abhängiger Variablen, genutzt werden, eine wichtige Ergänzung zum bisher vom Arbeitskreis Betriebswirtschaft verwendeten Gruppierungskriterium Betriebseinkommen/AK zu entwickeln, und damit sowohl die Abhängigkeit der Gruppierung der Betriebe von der gewählten Gruppierungskennzahl, als auch die in der Gruppierung am stärksten auffälligen Kennzahlen hervorheben.*

4.2.6 Klärung von multivariaten Beziehungsgefügen - graphische Modelle

Diskrete, nicht gerichtete, graphische Modelle werden in dieser Arbeit als einziges Verfahren eingesetzt um Variablenbeziehungen zu untersuchen. Die Klassenbildung erfolgt mit der bereits unter 4.2.2 aufgeführten Begründung. An zwei Ansätzen wird demonstriert, daß durch dieses Vorgehen wichtige Beziehungen und Zusammenhänge unter dem Aspekt der bedingten Unabhängigkeit aufgedeckt und graphisch dargestellt werden. In der vorliegenden Arbeit wird jedoch das volle Potential, daß in graphischen Modellen steckt noch nicht ausgeschöpft. Methodische Weiterentwicklungen bieten neue Möglichkeiten, die den hier verwendeten Beispielsdaten sogar noch angemessener sind (siehe 4.3). *Die Arbeit macht jedoch deutlich, daß graphische Modelle ein wirkungsvolles Instrumentarium bieten, um Zusammenhänge zwischen Merkmalen unter dem Aspekt der bedingten Unabhängigkeit zu untersuchen und darzustellen, und damit geeignet sind, Beziehungsgefüge multivariater Datensätze zu untersuchen.*

4.3 Kritik und Ausblick

Zum Abschluß ist nun zu überprüfen, ob die Zielsetzungen, die in der Einführung angesprochen werden, durch die Arbeit abgedeckt werden. Die vier Hypothesen zur explorativen Datenanalyse können bestätigt werden. Das eingesetzte Methodenspektrum erlaubt die Erstellung sinnvoller, graphischer Repräsentationen der vorliegenden Daten. Wirkungszusammenhänge werden zwar nicht letztlich geklärt, aber eine Vielzahl von Hypothesen kann aufgestellt und die Diskussion um Zusammenhänge angeregt und intensiviert werden. Strukturen werden sichtbar, wenn auch nicht, datenbedingt, in dem Umfang, daß völlig unbekannte Tatsachen zu Tage gefördert werden. Schließlich wird der Überprüfung von Annahmen und Resultaten auf verschiedensten Wegen Aufmerksamkeit geschenkt und somit voreiligen Schlüssen vorgebeugt. Darstellung und Umsetzung der verwendeten Methodik, zum Teil in eigenen Genstat Codes, unterstützen die gartenbauliche Beratung, wenn in der Zukunft diese Methoden eingesetzt werden sollen. Wenn die Arbeit möglicherweise auch nicht wirklich inhaltlich neue Erkenntnisse zu den verwendeten Beispielsdaten liefern kann, so gelingt doch in jedem Fall eine dichte und intensive Darstellung. Im Rahmen eines zuvor festgelegten Methodenspektrums wird eine stark interaktive Untersuchung der vorliegenden Daten durchgeführt, die schrittweise möglichst viele Problembereiche der Daten betrachtet und durchaus systematisch (wenn auch nicht schematisch) nach Erklärungsmustern und Auffälligkeiten sucht.

Demgegenüber stehen einige Defizite, die nicht unausgesprochen bleiben sollen:

1. Nicht alle einzelnen Analysen werden konsequent nach einem einheitlichen Schema durchgeführt. So wird zum Beispiel betont, wie wichtig die Überprüfung der Anzahl der ‘wesentlichen’ Dimensionen in der Hauptkomponentenanalyse ist, eine derartige Prüfung wird aber nicht in allen Fällen mit allen zur Verfügung stehenden Methoden durchgeführt. In der Regel wird versucht, einzelne, methodisch interessante Aspekte immer nur einmal darzustellen und nicht in jedem Abschnitt erneut zu wiederholen, sondern sich dann auf andere Schwerpunkte zu konzentrieren. Manch methodisch sinnvolles Vorgehen wird daher nur an einer Stelle durchgeführt, obwohl es auch an einer anderen Stelle angebracht wäre.
2. Die Frage nach der am besten für die Darstellung der Daten geeigneten Methode wird weder gestellt noch beantwortet. Dies vor allem darum, weil es eine Antwort aus Sicht des Verfassers nicht geben kann. Es mag sein, daß einige Darstellungsformen einen höheren Informationsgehalt haben als andere; ob dadurch aber auch per se die Mitteilung wesentlicher Inhalte eher gelingt bleibt einer weiteren Untersuchung vorbehalten. Darin wäre zu prüfen, wie unterschiedliche Personengruppen auf die Darstellung ein und derselben Dateninformation in unterschiedlichen Darstellungsformen reagieren, wie am schnellsten die vorhandenen Informationen erfaßt und wie gleichzeitig der Anteil an Desinformation minimiert werden können.

3. Das Methodenspektrum ist nicht vollständig und die Methodenwahl nur ansatzweise ausreichend inhaltlich begründet. Es ist allerdings so, daß gerade in der explorativen Phase einer Untersuchung, der Raum für das Ausprobieren verschiedener Methoden gegeben sein muß, um die vorliegenden Daten aus verschiedenen Blickwinkeln kennenlernen zu können. Die Methodenwahl richtet sich dann im wesentlichen danach, ob die Datenstruktur eine sinnvolle Analyse und Darstellung mit der entsprechenden Methode zuläßt. In dieser Arbeit wäre das Bemühen die Einschränkung auf das gewählte Methodenspektrum anders zu begründen, unehrlich. Sobald einzelne Ergebnisse der vorliegenden Untersuchungen jedoch außerhalb ihres Kontextes, das heißt außerhalb ihres Beitrags zur Datenexploration, diskutiert werden sollen, wird eine stringente Begründung, warum gerade ein spezielles Ergebnis einer speziellen Methodik verwendet wird, erforderlich. Aus der vorliegenden Arbeit ist eine derartige Begründung nicht zu entnehmen. Sie zeigt vielmehr, daß wenig perfekte Daten auch nur zu wenig perfekten Schlüssen führen können und es oft mindestes zwei ‘Wahrheiten’ zu ein und demselben Sachverhalt in den vorliegenden Daten gibt. Insofern mahnt die Arbeit dazu, sich nicht auf einzelne Ergebnisse zu verlassen, sondern immer wieder nach alternativen Darstellungs- und Analyseformen zu suchen. Da die Qualität der verwendeten Daten für die der Beratung vorliegenden Datenquellen sehr typisch ist, liegt die Stärke der Analyse und der Nutzen der Analysen für die Beratung auch weniger in der Gewinnung oder Ableitung eindeutiger Beratungsempfehlungen als vielmehr der Schaffung von mehr Transparenz und der Sensibilisierung für die Komplexität von in Daten enthaltenen Informationen.
4. Die inhaltlichen Ergebnisse mögen all jene nicht befriedigen, die sich aus der Sicht ihrer jeweiligen Disziplin speziell mit den in den Beispielsdaten angeschnittenen Themen auseinandersetzen oder auseinandergesetzt haben, weil möglicherweise nicht die Fragen beantwortet oder die Ansätze untersucht werden, die wichtiger oder untersuchungswürdiger zu sein scheinen. Es wird dadurch deutlich, wie sehr die (explorative) Datenanalyse ein miteinander von Datenanlytiker und Anwender braucht, um tatsächlich relevante Sachverhalte angemessen methodisch zu analysieren und inhaltlich zu interpretieren.

Aus datenanalytischer Sicht lassen sich die folgenden Aufgaben für weiterführende Ansätze in der explorativen Datenanalyse formulieren:

1. Der Einsatz weiterer, spezieller, interaktiver graphischer Software, die für diese Arbeit nicht zur Verfügung stand, ist zu überprüfen und zu bewerten. Es mag verwundern, daß interaktive graphische Werkzeuge, wie sie zum Beispiel in den Programmen SPSS/BMDP Diamond (BMDP, 1995) oder MANET (BIVAND, 1998) angeboten werden, in dieser Arbeit nicht in die Betrachtung miteinbezogen werden. SPSS/BMDP Diamond, das zum Teil eingesetzt wurde, konnte trotz verschiedener

Versuche keine vorzeigbaren Ergebnisse zu Tage fördern, da die Qualität der Ausdrucke aus diesem Programm von geringer Qualität und die zu produzierenden Graphiken vielfach wenig überzeugend sind (zum Beispiel Parametric Snake oder Fractal Foam). Zudem lebt eine Analyse mit einem interaktiven, graphischen Programm von der Bedienung und läßt sich nur schwer in Papierform darstellen. Der Verzicht auf derartige interaktive graphische Software bedeutet allerdings nicht, daß sie nicht auch in einer gelungenen Implementation durchaus informativ sein könnte. Übermäßige Erwartungen, wie sie UNWIN, 1992, oder THEUS, 1996, äußern, scheinen aber nicht angebracht. Einzelne überzeugende Beispiele (siehe zum Beispiel UNWIN, 1992) sollten nicht darüber hinwegtäuschen, daß auch diese Medien nicht nur durch die Möglichlichkeiten, die die Rechner heutzutage bieten, begrenzt sind, sondern auch dadurch, was der Nutzer gleichzeitig an Informationen von einer Bildschirmseite erfassen kann. Die Erfahrung mit den in dieser Arbeit eingesetzten Methoden zeigt, das diese Limitationen nicht übersehen werden sollten.

2. Lineare Strukturgleichungsmodelle werden ebenfalls in dieser Arbeit nicht eingesetzt; sie bieten aber grundsätzlich die Möglichkeit komplexe Beziehungszusammenhänge zu untersuchen und darzustellen. Allerdings wären zunächst inhaltlich begründete Modelle zu spezifizieren, die dann durch die Strukturgleichungsmodelle auf ihre Angemessenheit hin untersucht werden sollten. Mit Hilfe von Bootstrapping-Ansätzen ließen sich die starken Modellannahmen bezüglich der Multinormalverteilung abmildern (SMALLWATERS, 1997). Inwieweit aber die vorliegenden Beispielsdaten (und die Mehrzahl der der gartenbaulichen Beratung vorliegenden Daten besitzen in etwa dieselbe Qualität) ein derart stark konfirmatorisch geprägtes Verfahren rechtfertigen, ist zweifelhaft. Neben der genauen a priori Spezifizierung möglicher Modelle, stellt sich somit gleichzeitig die Forderung nach der Erschließung repräsentativer und aussagefähiger Datengrundlagen für den Gartenbau.
3. Das in graphischen Modellen steckende Potential zur Analyse und Darstellung multivariater Beziehungsgefüge wird durch die vorliegende Arbeit nicht ausgeschöpft. Es bleibt zu klären, ob nicht kontinuierliche und gemischte graphische Modelle nicht doch bei entsprechender Transformation einzelner Variablen eingesetzt werden sollten. Darüberhinaus werden in den letzten Jahren verstärkt Entwicklungen auf dem Gebiet der, den graphischen Modellen verwandten, Markov Chain Monte Carlo-Methoden (MCMC) diskutiert, die für komplexe Beziehungszusammenhänge bei wenig perfekten Daten sinnvoll eingesetzt werden können (siehe zum Beispiel BEST et al., 1996, GOLDSTEIN & SPIEGELHALTER, 1996).

5 Zusammenfassung

Ausgangspunkt der vorliegenden Arbeit ist die Suche der gartenbaulichen Beratung nach Visualisierungsmöglichkeiten umfangreicher gartenbaulicher Datensätze, die einerseits zu einer graphischen Zusammenfassung der in den Daten enthaltenen Informationen dienen und die andererseits auf interaktivem Weg Möglichkeiten der graphischen Analyse von Erhebungsdaten liefern.

Die weitgehende Freiheit von Modellannahmen, der überwiegend deskriptive Charakter der Untersuchungen, das interaktive, schrittweise Vorgehen in der Auswertung, und die Betonung graphischer Elemente kennzeichnet die Arbeit als Beitrag zur explorativen Datenanalyse.

Das ausgewählte Methodenspektrum, das ausführlich besprochen wird, schließt Verfahren der Dimensionserniedrigung (Hauptkomponentenanalyse, Korrespondenzanalyse und mehrdimensionale Skalierung) und darauf aufbauende Biplots, die Analyse gruppierter Daten (Prokrustes-Rotation und Gruppenanalysemodelle in der Hauptkomponentenanalyse), Linienverbände (Liniendiagramme der formalen Begriffsanalyse, Baumdiagramme und graphische Modelle), sowie ergänzende graphische Verfahren, wie zum Beispiel Trellis-Displays, ein.

Beispielhaft werden eine betriebsbegleitende Untersuchung mit Cyclamen aus der Beratungspraxis der Landwirtschaftskammer Westfalen-Lippe und die Kennzahlen der Jahre 1992 bis 1994 der Topfpflanzenbetriebe des Arbeitskreises für Betriebswirtschaft im Gartenbau aus Hannover analysiert. Neben einer Vielzahl informativer Einzelergebnisse, zeigt die Arbeit auch auf, daß die qualitativ relativ schlechten Datengrundlagen nur selten eindeutige Schlußfolgerungen zulassen. Sie sensibilisiert also in diesem Bereich für die Problematik, die der explorativen Analyse wenig perfekter Daten innewohnt.

Als besonders sinnvolle Hilfsmittel in der graphischen Analyse erweisen sich Biplots, hierarchische Liniendiagramme und Trellis-Displays. Die Segmentierung einer Vielzahl von Objekten in einzelne Gruppen wird durch Klassifikations- und Regressionsbäume vor allem unter dem Gesichtspunkt der Visualisierung gut gelöst, da den entstehenden Baumstrukturen auch die die Segmente bestimmenden Variablen visuell entnommen werden können. Diskrete graphische Modelle bieten schließlich einen guten Ansatzpunkt zur Analyse von multivariaten Beziehungszusammenhängen.

Einzelne, nicht in der statistischen Standardsoftware vorhandene Prozeduren sind in eigens erstellten Programmcodes zusammengefaßt und können mit dem Programm Genstat genutzt werden.

Literaturverzeichnis

Anderson, T. W., 1963 Asymptotic Theory for Principal Component Analysis Annals of Mathematical Statistics 1963 *43* Seite 122 - 148

Andrews, D.E., 1972 Plots of High Dimensional Data Biometrics 1972 *28* Seite 125 - 136

Andrews, D.F., Gnanadesikan, R., Warner, J.L., 1971 Transformations of Multivariate Data Biometrics 1971 *27* Seite 825 - 840

Arabie, P., 1973 Concerning Monte Carlo Evaluations of Nonmetric Multidimensional Scaling Algorithms Psychometrika 1973 *38* Seite 607 - 608

Arabie, P., 1978a Random versus Rational Strategies for Initial Configurations in Nonmetric Multidimensional Scaling Psychometrika 1978 *43* Seite 111 - 113

Arabie, P., 1978b The Difference Between „Several“ and „Single“. A Reply to Spence and Young Psychometrika 1978 *43* Seite 119

Arnold, G. M. & Collins, A. J., 1993 Interpretation of Transformed Axes in Multivariate Analysis Applied Statistics 1993 *42* Seite 381 - 400

Bahnmüller, H., 1997 Beängstigende Einbrüche Deutscher Gartenbau 1997 *18* Seite 1050 - 1052

Bahnmüller, H., 1998 Orientierungsdaten 1998 Deutscher Gartenbau 1998 *17* Seite 9 - 11

Banfield, J.D. & Raftery, A.E., 1992 Model-Based Gaussian and Non-Gaussian Clustering Biometrics 1992 *49* Seite 803 - 822

Bartholomew, D. J., 1980 Factor Analysis for Categorical Data Journal of the Royal Statistical Society Series B 1980 *42* Seite 293 - 321

Bartholomew, D. J., 1984 The Foundations of Factor Analysis Biometrika 1984 *76* Seite 221 - 232

Bartholomew, D. J., 1985 Foundations of Factor Analysis: Some Practical Implications British Journal of Mathematical and Statistical Psychology 1985 *38* Seite 1 - 10

Bartlett, M. S., 1950 Tests of Significance in Factor Analysis British Journal of Psychology Statistics Section 1950 *3* Seite 77 - 85

Bartlett, M. S., 1954 A Note on Multiplying Factors for Various Chi-Squared Approximations Journal of the Royal Statistical Society Series B 1954 *16* Seite 296 - 298

Bauer, H. & Teutter, D., 1990 Identification of Pelargonium Genotypes by Phenolic ‘Fingerprints’ II. Cultivar Identification by HLPC Analysis of Leaf Phenols Combined with Discriminant Analysis Gartenbauwissenschaft 1990 *55* Seite 187 - 191

Beale, E. M. L. & Little, R. J. A., 1975 Missing Values in Multivariate Analysis Journal of the Royal Statistical Society Series B 1975 *37* Seite 129 - 145

Best, N.G., Spiegelhalter, D.J., Thomas, A., Brayne, C.E.G., 1996 Baysian Analysis of Realistically Complex Models Journal of the Royal Statistical Society Series A 1996 *159* Seite 323 - 342

Beyl, C.A., Ghale, G., Zhang, L., 1995 Characteristics of Hardwood Cuttings Influence Rooting of Actinidia arguta (Siebold & Zucc.) Planch. HortScience 1995 *30* Seite 973 - 976

Bivand, R.S., 1998 Software and Software Design Issues in the Exploration of Local Dependence The Statistician 1998 *47* Seite 499 - 508

Bock, H.H., 1985 On Some Significance Tests in Cluster Analysis Journal of Classification 2 *1985* Seite 77 -108

Box, G.E.P., Hunter, W.G., MacGregor, J.F., Erjavac, J., 1973 Some Problems Associated with the Analysis of Multiresponse Data Technometrics 1973 *15* Seite 33 - 51

Brady, H. E., 1985 Statistical Consistency and Hypothesis Testing for Nonmetric Multidimensional Scaling Psychometrika 1985 *50* Seite 503 - 537

Burt, C., 1950 The Factorial Analysis of Qualitative Data British Journal of Mathamatical and Statistical Psychology, 1950 *3* Seite 166 - 185

Campbell, N.A., 1980 Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation Applied Statistics 1980 *29* Seite 231 - 237

Carrol, J. D., 1953 An Analytical Solution for Approximating Simple Structure in Factor Analysis Psychometrika 1953 *18* Seite 23 - 38

Carrol, J. D., Chang, J. J., 1970 Analysis of Individual Differences in Multidimensional Scaling Via an N-Way Generalisation of the 'Eckart-Young' Decomposition Psychometrika 1970 *35* Seite 283 - 319

Cattel, R. B., 1966 The Scree Test for Number of Factors Multivariate Behavioural Research 1966 *1* Seite 245 - 276

Chatfield, C., 1997 Data Mining RSS News 1997 *25* Seite 1 - 2

Corsten, L.C.A. & Gabriel, K.R., 1976 Graphical Exploration in Comparing Variance Matrices Biometrics 1976 *32* Seite 851 - 863

Cruz-Castillo, J.G., Ganeshanandam, R.R., MacKay, B.R., Lawes, G.S., Lawoko, C.R.O., Woolley, D.J., 1994 Applications of Canonical Discriminant Analysis in Horticultural Research HortScience 1994 *29* Seite 1115 - 1119

Degani, C., Beiles, A., El-Batsri, R., Goren, M., Gazit, S., 1995 Identifying Lychee Cultivars by Isozyme Analysis Journal of the American Society of Horticultural Science 1995 *120* Seite 307 - 312

Dever, M.C., MacDonald, R.A., Cliff, M.A., Lane, W.D., 1996 Sensory Evaluation of Sweet Cherry Cultivars HortScience 1996 *31* 150 - 153

Eastment, H.T. & Krzanowski, W.J., 1982 Cross-validatory Choice of the Number of Components from a Principal Components Analysis Technometrics 1982 *24* Seite 73 - 77

Edwards, D. & Havránek, T., 1985 A Fast Procedure for Model Search in Multidimensional Contingency Tables Biometrika 1985 *72* Seite 339 - 351

Edwards, D. & Havránek, T., 1987 A Fast Model Selection Procedure for Large Families of Models Journal of the American Statistical Association 1987 *82* Seite 205 - 213

Everitt, B. S., 1979 Unresolved Problems in Cluster Analysis Biometrics 1979 *35* 169 - 181

Fabbri, A., Hormazy, J.J., Polito, V.S., 1995 Random Amplified Polymorphic DNA Analysis of Olive (Olea europaea L.) Cultivars Journal of the American Society of Horticultural Science 1995 *120* Seite 538 - 542

Fernandez, A., Schutzki, R.E., Hancock, J.F., 1996 Isozyme and Morphological Variation in a Cornus florida L. Provenance Plantation Representing Geographically Diverse Populations Journal of the American Society of Horticultural Science 1996 *121* Seite 225 - 230

Finney, D.J., 1956 Multivariate Analysis and Agricultural Experiments Biometrics 1956 *12* Seite 67 - 71

Fischer, M., 1993 Anwendung der Diskriminanzanalyse in der Apfelunterlagen-Selektion Gartenbauwissenschaft 1993 *58* Seite 137 - 143

Flury, B. N., 1984 Common Principal Components in K Groups Journal of the American Statistical Association 1984 *79* Seite 892 - 898

Gabriel, K. R., 1971 The Biplot Graphic Display of Matrices with Application to Principal Component Analysis Biometrika 1971 *58* Seite 453 - 467

Goldstein, H. & Spiegelhalter, D.J., 1996 Statistical Aspects of Institutional Performance: League Tables and Their Limitations Journal of the Royal Statistical Society Series A 1996 *159* Seite 385 - 444

Good, I. J., 1969 Some Applications of the Singular Value Decomposition of A Matrix Technometrics 1969 *11* Seite 823 - 831

Gower, J. C. & Harding, S., 1988 Nonlinear Biplots Biometrika 1988 *75* Seite 445 - 455

Gower, J. C. & Legendre, P., 1986 Metric and Euklidean Properties of Dissimilarity Coefficients Journal of Classification 1986 *3* Seite 5 - 48

Gower, J. C., 1966 Some Distance Properties of Latent Root and Related Methods Used in Multivariate Analysis Biometrika 1966 *53* Seite 325 - 338

Gower, J. C., 1971 A General Coefficient of Similarity and Some of Its Properties Biometrics 1971 *27* Seite 857 - 872

Gower, J. C., 1975 Generalized Procentes Analysis Psychometrika 1975 *40* Seite 33 - 51

Gower, J.C. & Ross, G.J.S., 1969 Minimum Spanning trees and Single Linkage Cluster Analysis Applied Statistics 1969 *18* Seite 54 - 64

Greenacre, M. J., 1988 Correspondence Analysis of Multivariate Categorical Data by Weighted Least Squares Biometrika 1988 *3* Seite 457 - 467

Greenacre, M. J., 1990 Some Limitations of Multiple Correspondence Analysis Computational Statistics Quarterly 1990 *3* Seite 249 - 256

Greenacre, M. J., 1991 Interpreting Multiple Correspondence Analysis Applied Stochastic Models and Data Analysis 1991 *7* Seite 195 - 210

Guttman, L., 1954 Some Necessary Conditions for Common-Factor Analysis Psychometrika 1954 *19* Seite 149 - 161

Harris, P. Testing for Variance Homogeneity of Correlated Variables Biometrika 1985 *72* Seite 103 - 107

Hawkins, D.M., 1974 The Detection of Errors in Multivariate Data Using Principal Components Journal of the American Statistical Association 1974 *69* Seite 340 - 344

Hill, M. O., 1974 Correspondence Analysis: A Neglected Multivariate Method Applied Statistics 1974 *23* Seite 340 - 354

Hills, M., 1977 Book Review Applied Statistics 1977 *26* Seite 339 - 340

Horn, J.L., 1965 Psychometrika 1965 30 *Seite 179 - 185*

Joliffe, I. T., 1972 Discarding Variables in Principal Component Analysis I: Artificial Data Applied Statistics 1972 *21* Seite 160 - 173

Joliffe, I. T., 1973 Discarding Variables in Principal Component Analysis II: Real Data Applied Statistics 1973 *22* Seite 21 - 31

Jones, C. R., 1983 A Note of the Use of Directional Statistics in Weigthted Euclidean Multidimensional Scaling Models Psychometrika 1983 *48* Seite 473 - 476

Kaiser, H. F., 1959 Computer Program for Varimax Rotation in Factor Analysis Journal of Educational and Psychological Measurement 1959 *19* Seite 413 - 420

Kass, G.V., 1980 An Exploratory Technique for Investigating Large Quantities of Categorical Data Applied Statistics 1980 *29* Seite 119 - 127

Keramidas, E. M., Devlin, S. J., Ganadesikan, R., 1987 A Graphical Procedure for Comparing the Principal Components of Several Covariance Matrices Communications in Statistics; Simulation and Computation 1987 *16* Seite 161 - 191

Klahr, D., 1969 A Monte Carlo Investigation of the Statistical Significance of Kruskals Nonmetric Scaling Procedure Psychometrika 1969 *34* Seite 319 - 330

Koziol, J.A., 1986 Assessing Multivariate Normality: A Compendium Communications in Statistics: Theory and Methods 1986 *15* Seite 2763 - 2783

Kruskal, J., 1964a Multidimensional Scaling by Optimising Goodness of Fit to A Nonmetric Hypothesis Psychometrika 1964 *29* Seite 1 - 28

Kruskal, J., 1964b Nonmetric Multidimensional Scaling: A Numerical Method Psychometrika 1964 *29* Seite 115 - 129

Krzanowski, W. J., 1979 Between Groups Comparison of Principal Components Journal of the American Statistical Association 1979 *74* Seite 703 - 707 (Korrektur in 76, 1022)

Krzanowski, W. J., 1984 Principal Components Analysis in the Presence of Group Structure Applied Statistics 1984 *33* Seite 164 - 168

Krzanowski, W.J. & Lai, Y.T., 1992 A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering Biometrics 1988 *44* Seite 23 - 34

Krzanowski, W.J., 1988b Missing Value Imputation in Multivariate data Using the Singular Value Decomposition of a Matrix Biometrical Letters 1988 *25* 31 -39

Krzanowski, W.J., 1993 Attribute Selection in Correspondence Analysis of Incidence Matrices Applied Statistics 1993 *42* Seite 529 - 541

Lawley, D. N., 1956 Tests of Significance for the Latent Roots of Covariance and Correlation Matrices Biometrika 1956 *43* Seite 128 - 136

Levine, D. M., 1978 A Monte Carlo Study of Kruskals Variance Based Measure on Stress Psychometrika 1978 *43* Seite 307 - 315

Loh, W-Y. & Vanichsetakul, N., 1988 Tree-Structured Classification Via Generalized Discriminant Analysis Journal of the American Statistical Association 1988 *83* Seite 715 - 728

Machado, S.G., 1983 Two Statistics for Testing Multivariate Normality Biometrika 1983 *70* Seite 713 - 718

Malkovich, J.F. & Afifi, A.A., 1973 On Tests for Multivariate Normality Journal of the American Statistical Association 1973 *68* Seite 176 - 179

McArdle, J. J., 1990 Principles versus Principals of Structural Factor Analysis Multivariate Behavioural Research 1990 *25* Seite 81 - 87

McDonald, R. P., 1985 Comments on D. J. Batholomew: Foundations of Factor Analysis: Some Practical Implications British Journal Mathematical & Statistical Psychology 1985 *38* Seite 134 - 137

Milligan, G.W. & Cooper, M.C., 1985 An Examination of Procedures for Determining the Number of Clusters in a Data Set Psychometrika 1985 *50* Seite 159 - 179

Mulaik, S. A., 1990 Blurring the Distinction Between Component Analysis and Common Factor Analysis Multivariate Behavioural Reseach 1990 *25* Seite 53 - 59

Nienhuis, J., Tivang, J., Skroch, P., dos Santos, J.B., 1995 Genetic Relationships Among Cultivars and Landraces of Lima Bean (Phaseolus lunatus L.) As Measured by RAPD Markers Journal of the American Society of Horticultural Science 1995 *120* Seite 300 - 306

Novy, R.G., Vorsa, N., Patten, K., 1996 Identifying Genotypic Heterogeneity in 'McFarlin' Cranberry: A Randomly-amplified Polymorphic DNA (PAPD) and Phenotypic Analysis Journal of the American Society of Horticultural Science 1996 *121* Seite 210 - 215

PANEL (on Discriminant Analysis, Classification, and Clustering), 1989 Discriminant Analysis and Clustering Statistical Science 1989 *4* Seite 34 - 69

Parent, J.E., Isfan, D., Tremblay, N., Karam, A., 1994 Multivariate Nutrient Diagnosis of the Carrot Crop Journal of the American Society of Horticultural Science 1994 *119* Seite 420 - 426

Pearce, S.C. & Holland, D.A., 1960 Some Applications of Multivariate Methods in Botany Applied Statistics 1960 *9* Seite 1 - 7

Pereira-Lorenzo, S., Fernandez-López, J., Moreno-González, J., 1996a Variability and Grouping of Northwestern Spanish Chestnut Cultivars: I. Morphological Traits Journal of the American Society of Horticultural Science 1996 *121* Seite 183 - 189

Pereira-Lorenzo, S., Fernandez-López, J., Moreno-González, J., 1996b Variability and Grouping of Northwestern Spanish Chestnut Cultivars: II. Isoenzyme Traits Journal of the American Society of Horticultural Science 1996 *121* Seite 190 - 197

Plotto, A., Azarenko, A.N., McDaniel, M.R., Crocket, P.W., Mattheis, J.P., 1997 Eating Quality of 'Gala' and 'Fuji' Apples from Multiple Harvests and Storage Durations HortScience 1997 *32* Seite 903 - 908

Ramsey, J. O., 1977 Maximum Likelihood Estination in Multidimensional Scaling: Theory and Applications in the behavioural sciences Psychometrika 1977 *42* Seite 241 - 266

Ramsey, J. O., 1978 Confidence Regions for Multidimensional Scaling Analysis Psychometrika 1978 *43* Seite 145 - 160

Ramsey, J. O., 1980 Some Small Sample Results for Maximum Likelihood Estimation in Multidimensional Scaling Psychometrika 1980 *45* Seite 139 - 144

Ramsey, J. O., 1982 Some Statistical Approaches to Multidimensional Scaling Data Journal of the Royal Statistical Society Series A 1982 *145* Seite 285 - 312

Rath, T., 1996 Klassifikation und Identifikation gartenbaulicher Objekte mit künstlichen neuronalen Netzwerken Gartenbauwissenschaft 1996 *61* Seite 153 - 159

Ren, J., McFerson, J.R., Li, R., Kresovich, S., Lamboy, W.F., 1995 Identities and Relationships Among Chinese Vegetable Brassicas as Determind by Random Amplified Polymorphic DNA Markers Journal of the American Society of Horticultural Science 1995 *120* Seite 548 - 555

Royston, J.P., 1983 Some Techniques for Assessing Normality Based on the Shapiro-Wilk W Applied Statistics 1983 *32* Seite 121 - 133

Rumayor-Rodríguez, A., 1995 Multiple Regression Models for the Analysis of Potential Cultivation Areas for Japanese Plums HortScience 1995 *30* Seite 605 - 610

Schönemann, P. H., 1990 Facts, Fictions, and Common Sence About Factors and Components Multivariate Behavioural Reseach 1990 *25* Seite 42 - 51

Schott, J. R., 1988 Testing the Quality of the Smallest Latent Roots of a Correlation Matrix Biometrika 1988 *75* Seite 794 - 796

Shepard, R. N., 1962a The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function I Psychometrika 1962 *27* Seite 125 - 139

Shepard, R. N., 1962b The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function II Psychometrika 1962 *27* Seite 219 - 249

Sibson, R., 1979 Studies in the Robustness of Multidimensional Scaling: Pertubational Analysis of Classial Scaling Journal of the Royal Statistical Society Series B 1979 *41* Seite 217 - 229

Simpson, C.H., 1951 The Interpretation of Interaction in Contingency Tables Journal of the Royal Statistical Society Series B 1951 *13* Seite 238 - 241

Small, N.J.H., 1980 Marginal Skewness and Kurtosis in Testing Multivariate Normality Applied Statistics 1980 *29* Seite 85 - 87

Spence, I. & Lewandowsky, S., 1989 Robust Multidimensional Scaling Psychometrika 1989 *54* Seite 501 - 513

Spence, I., 1972 Monte Carlo Evaluation of Three Nonmetric Multidimensional Scaling Algorithms Psychometrika 37 *1972* Seite 461 - 486

Spence, I., 1979 A Simple Approximation for Random Ranking Stress Values Multivariate Behavioural Research 1979 *14* Seite 355 - 365

Spence, I., Young, F. W., 1978 Monte Carlo Studies in Nonmetric Scaling Psychometrika 1978 *43* Seite 115 - 117

Steel, R.G.D., 1955 An Analysis of Perennial Crop Data Biometrics 1955 *11* Seite 201 - 212

Steinbacher, F., Wehrpfenning, M., Hauser, B., 1995 Homogene Bestände durch kürzere Gießrhythmen Deutscher Gartenbau 1995 *30* Seite 1790 - 1793

Tivang, J., Skorch, P.W., Nienhuis, J., De Vos, N., 1996 Randomly Amplified Polymorphic DNA (RAPD) Variation among and within Artichoke (Cynara scolymus L.) Cultivars and Breeding Populations Journal of the American Society of Horticultural Science 1996 *121* Seite 783 - 788

Unwin, A., 1992 How Interactive Graphics Will Revolutionize Statistical Practice The Statistician 41 *1992* Seite 365 - 369

van der Heijden, P. G. M. de Falguerolles, A., de Leeuws, J., 1989 A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Models Applied Statistics 1989 *38* Seite 249 - 292

van der Heijden, P. G. M., de Leeuws, J., 1985 Correspondence Analysis Used Complementory to Loglinear Analysis Psychometrika 1985 *50* Seite 429 - 441

Velicer, W. F. & Jackson, D. N., 1990b Component Analysis versus Common Factor Analysis: Some Further Observations Multivariate Behavioural Research 1990 *25* Seite 97 - 114

Velicer, W. F., & Jackson, D. N., 1990a Component Analysis Versus Common Factor Analysis: Some Issues in Selecting an Appropriate Procedure Multivariate Behavioural Reseach 25 *1990* Seite 1 - 28

Velicer, W.F., 1976 Determining the Number of Components from the Matrix of Partial Correlations Psychometrika 1976 *41* Seite 321 - 327

Wegman, E., 1990 Hyperdimensional Data Analysis Using Parallel Coordinates Journal of the American Statistical Association 1990 *85* Seite 664 - 675

AKBWL (Arbeitskreis Betriebs-wirtschaft im Gartenbau e.V.), 1996 Kennzahlen für den Betriebsvergleich - Heft 39 - Eigendruck Hannover 1. Auflage 1996

Bacher, J., 1994 Clusteranalyse R. Oldenbourg München 1. Auflage 1994

Benzécri, J. P., 1973 Analyse des Correspondance Durod Paris 1. Auflage 1973

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984 Classification and Regression Trees Wadsworth Belmont 1. Auflage 1984

Chatfield, C., 1995 Problem Solving: A Statisticians Guide Chapman & Hall London 2. Auflage 1995

Chatfield, C. & Collins, A. J., 1980 Introduction to Multivariate Analysis Chapman & Hall London 1. Auflage 1980

Cleveland, W.S., 1993 Visualising Data Hobart Press Summit 1. Auflage 1993

Cox, D.R. & Wermuth, N., 1996 Multivariate Dependencies - Models Analysis and Interpretation Chapman & Hall London 1. Auflage 1996

Deichsel, G. & Trampisch, H. J., 1985 Clusteranalyse und Diskriminanzanalyse Gustav Fischer Verlag Stuttgart 1. Auflage 1985

Edwards, D., 1995 Introduction to Graphical Modelling Springer Heidelberg 1. Auflage 1995

Everitt, B. S., 1978 Graphical Techniques for Multivariate Data Heineman London 1. Auflage 1978

Everitt, B. S., 1980 Cluster Analysis Heineman London 2. Auflage 1980

Fienberg, S.E., 1980 The Analysis of Cross-Classified Categorical Data MIT Press London 2. Auflage 1980

Flury, B. N., 1988 Common Principal Components and Related Multivariate Models Wiley New York 1. Auflage 1988

Flury, B. N. & Riedwyl, H., 1988 Applied Multivariate Statistical Methods Chapman & Hall London 1. Auflage 1988

Genstat Committee, 1993 Genstat Release 3 Reference Manual Clarendon Press Oxford 1. Auflage 1993

Gifi, A., 1990 Nonlinear Multivariate Analysis Wiley New York 1. Auflage 1990

Gnanadesikan, R., 1977 Methods for Statistical Data Analysis of Multivariate Observations Wiley New York 1. Auflage 1977

Gordon, A. R., 1981 Classification Chapman & Hall London 1. Auflage 1981

Gower, J. C. & Hand, D. J., 1996 Biplots Chapman & Hall London 1. Auflage 1996

Greenacre, M. J., 1984 Theory and Applications of Correspondence Analysis Academic Press London 1. Auflage 1984

Greenacre, M. J., 1993 Correspondence Analysis in Practice Academic Press London 1. Auflage 1993

Hand, D. & Crowder, M., 1996 Practical Longitudinal Data Analysis Chapman & Hall London 1. Auflage 1996

Harman, H. H., 1976 Modern Factor Analysis Universitiy of Chicago Press Chicago 3. Auf-lage 1976

Hawkins, D.M., 1980 Identification of Outliers Chapman & Hall London 1. Auflage 1980

Jackson, J. E., 1991 A User's Guide to Principal Components Wiley New York 1. Auflage 1991

Jambu, M., 1992 Explorative Datenanalyse Gustav Fischer Verlag Stuttgart 1. Auflage 1992

Jardine, F. & Sibson, R., 1972 Mathematical Taxonomy Wiley New York 1. Auflage 1972

Jokiel, A. & Hockwien, J., 1994 Cyclamen Praxisversuch Landwirtschaftskammer Westfalen Lippe Eigendruck Münster 1. Auflage 1994

Joliffe, I. T., 1986 Principal Component Analysis Springer Verlag Heidelberg 1. Auflage 1986

Jöreskog, K. G. & Sörenbom, D., 1993 Lisrel 8: Structural Equation Modelling with the SIMPLIS Commend Language Lawrence Erlbaum Hillsdale 1. Auflage 1993

Kaufman, L. & Rousseeuw, P.J., 1990 Finding Groups in Data: An Introduction to Cluster Analysis Wiley New York 1. Auflage 1990

Krzanowski, W. J., 1988a Principles of Multivariate Analysis: A User's Perspective Clarendon Press Oxford 1. Auflage 1988

Krzanowski, W.J. (Hrsg.), 1995 Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1. Auflage 1995

Krzanowski, W.J. & Marriott, F.H.C., 1994 Multivariate Analysis Part 1: Distributions, Ordination and Inference Arnold London 1. Auflage 1994

Krzanowski, W.J. & Marriott, F.H.C., 1995 Multivariate Analysis Part 2: Classification, Covariance Structures and Repeated Measurements Arnold London 1. Auflage 1995

Lauritzen, S.L., 1996 Graphical Models Clarendon Press Oxford 1. Auflage 1996

Lazarsfeld, P. F. & Henry, W. W., 1968 Latent Structure Analysis Houghton-Miffiln Boston 1. Auflage 1968

Manly, B. F. J., 1986 Multivariate Statistical Methods: A Primer Chapman & Hall London 1. Auflage 1986

Marriott, F.H.C., 1974 The Interpretation of Multiple Observations Academic Press London 1. Auflage 1974

MathSoft, 1997 S-Plus 4. Guide to Statistics MathSoft Seattle 1. Auflage 1997

Morrison, D. F., 1990 Multivariate Statistical Methods Mc Graw-Hill New York 3. Auflage 1990

Nagel, M., Benner, A., Ostermann, R., Henschke, K., 1996 Graphische Datenanalyse Gustav Fischer Verlag Stuttgart 1. Auflage 1996

Pfeifer, A. & Schmidt, P., 1987 Lisrel: Die Analyse komplexer Strukturgleichungsmodelle Gustav Fischer Verlag Stuttgart 1. Auflage 1987

Rasch, D., Guiard, V., Nürnberg, G., 1992 Statistische Versuchsplanung Gustav Fischer Verlag Stuttgart 1. Auflage 1992

Schiffman, S. S., Reynolds, M. L., Young, F. W., 1981 Introduction to Multidimensional Scaling: Theory, Methods and Applications Academic Press New York 1. Auflage 1981

Schubö, W., Uehlinger, H.-M., Perleth, Ch., Schröger, E., Sierwald, W., 1991 SPSS-Handbuch der Programmversionen 4.0 und SPSS - X 3.0 Gustav Fischer Verlag Stuttgart 1. Auflage 1991

Seber, G.A.F., 1984 Multivariate Observations Wiley New York 1. Auflage 1984

Shao, J. & Tu, D., 1996 The Jackknife and Bootstrap Springer Heidelberg 1. Auflage 1996

SmallWaters, 1997 Amos User's Guide, Version 3.6 SmallWaters Chicago 1. Auflage 1997

Sokal, R.R., Rohlf, F.J. Biometry Freeman New York 2. Auflage 1981

Sprent, P., 1997 Data Driven Statistical Methods Chapman & Hall London 1. Auflage 1997

SPSS, 1993 SPSS for Windows CHAID, Release 6.0 SPSS Chicago 1. Auflage 1993

SPSS, 1994 SPSS Categories 6.1 SPSS Chicago 1. Auflage 1994

Stevens, S.S., 1951 Handbook of Experimental Psychology Wiley New York 1. Auflage 1951

Storck, H. & Bokelmann, W., 1995 Grundzüge der gartenbaulichen Betriebslehre Eugen Ulmer Stuttgart 1. Auflage 1995

Torgerson, W. S., 1958 Theory and methods of Scaling Wiley London 1. Auflage 1958

Whittaker, J., 1990 Graphical Models in Applied Multivariate Statistics Wiley New York 1. Auflage 1990

Young, F. W., 1987 Multidimensional Scaling: History: Theory and Applications Lawrence Erlbaum Hillsdale 1. Auflage 1987

Arabie, P. & Hubert, L.J., 1995 Clustering from the Perspective of Combinatorical Data Analysis in: Krzanowski, W. J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 1 - 13

Baráth, E., 1993 Anwendung multivariater Methoden in der Agrarforschung in: Meister, R. & Weiß, H. (Hrsg.) Kolloquium Statistische Methoden in der experimentellen Forschung Kurzfassung von Vorträgen der Wintersemester 1991/92 und 1992/93 Eigendruck Berlin 1993 1. Auflage Seite 115 - 125

Bock, H. H., 1992 Grundlegende Methoden der exploratorischen Datenanalyse in: Encke, H., Göller, J., Haux, R., Wernecke, R. D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Stuttgart 1992 1. Auflage Seite 15 - 42

Borovnik, M., 1992 Diskussion zu H.H. Bock: Grundlegende Methoden der exploratorischen Datenanalyse in: Enke, H., Gölles, J., Haux, R., Wernecke, K.-D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Verlag Stutgart 1992 1. Auflage Seite 43 - 46

Carrol, J. D., 1972 Individual Differences and Multidimensional Scaling in: Shepard, R. N., Romney, A. K., Nerlove, S. G. (Hrsg.) Multidimensional Scaling: Theory and Applications in the Behavioural Sciences, Volume I Seminar Press New York 1972 1. Auflage Seite 105 - 157

Gabriel, K. R., 1981 Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis in: Barnett, V. (Hrsg.) Interpreting Multivariate Data Wiley New York 1981 1. Auflage Seite 147 - 174

Gabriel, K. R., 1995a Biplot Display of Multivariate Categorical Data with Comments on Multiple Correspondence Analysis in: Krzanowski, W. J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 190 - 226

Gabriel, K. R., 1995b MANOVA Biplots for Two-Way Contingency Tables in: Krzanowski, W. J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 227 - 268

Gabriel, K. R. & Odoroff, C. L., 1986 Use of Three Dimensional Biplots for Diagnosis of Models in: Gaul, W. & Schrader, M. (Hrsg.) Classification as a Tool of Research Elsevier Science Publishers North Holland 1986 1. Auflage Seite 153 - 159

Gower, J. C., 1993 The Construction of Neighbour Regions in Two Dimensions for Prediction with Multilevel Categorical Variables in: Opitz, O., Lauser, B. & Klar, R. (Hrsg.) Information and Classification: Concepts-Methods-Applications: Proceedings 16th Annual Conference of the Gesellschaft für Klassifikation, Dortmund, April 1992 Springer Verlag Heidelberg 1993 1. Auflage Seite 174 - 189

Gower, J. C., 1995a Orthogonal and Projection Procentes Analysis in: Krzanowski, W. J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 113 - 134

Gower, J. C., 1995b A General Theory of Biplots in: Krzanowski, W. J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 283 - 303

Green, P.J., 1981 Peeling Bivariate Data in: Barnett, V. (Hrsg.) Interpreting Multivariate Data Wiley New York 1981 1. Auflage Seite 3-20

Greenacre, M. J., 1982 Practical Correspondence Analysis in: Barnett, V. (Hrsg.) Interpreting Multivariate Data Wiley New York 1981 1. Auflage Seite 119-146

Hass-Tschirschke, I., 1994 Cyclamen, Alpenveilchen in: Röber, R. (Hrsg.) Topfpflanzenkulturen Eugen Ulmer Verlag Stuttgart 1994 7. Auflage

Heiser, W.J. & Meulman, J.J., 1995 Nonlinear Methods for the Analysis of Homogeneity and Heterogeneity in: Krzanowski, W.J. (Hrsg.) Recent Advances in Descriptive Multivariate Analysis Clarendon Press Oxford 1995 1. Auflage Seite 51 - 89

Horn, W., 1996 Cyclamen L., Alpenveilchen in: Horn, W. (Hrsg.) Zierpflanzenbau Blackwell Berlin 1996 1. Auflage

Lengnink, K., 1993 Diagrams of Similarity Matrices in: Opitz, O., Lausen, B. Klar, R. (Hrsg.) Information and Classification Springer Heidelberg 1993 1. Auflage Seite 99 - 107

Meulman, J. J. & Heiser, W. J., 1993 Nonlinear Biplots for Nonlinear Mappings in: Opitz, O., Lauser, B. & Klar, R. (Hrsg.) Information and Classification: Concepts-Methods-Applications: Proceedings 16th annual conference of the Gesellschaft für Klassifikation, Dortmund, April 1992 Springer Verlag Heidelberg 1993 1. Auflage Seite 201 - 213

Nagel, M., Hothorn, L., Hartmann, P., 1992 Hochinteraktive Datenanalyse - Werkzeuge und Prinzipien in: Enke, H., Gölles, J., Haux, R., Wernecke, K.-D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Verlag Stutgart 1992 1. Auflage Seite 75 - 91

Ostermann, R. & Wolf-Ostermann, K., 1992 Moderne interaktive Graphik mit den klassischen Auswertungssystemen BMDP, SAS und SPSS unter Zuhilfenahme von ISP in: Enke, H., Gölles, J., Haux, R., Wernecke, K.-D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Verlag Stuttgart 1992 1. Auflage Seite 95 - 107

Rovan, J., 1994 Visualizing Solutions in More Than Two Dimensions in: Greenacre, M. & Blasius, J. (Hrsg.) Correspondence Analysis in the Social Sciences Academic Press London 1994 1. Auflage

Schiller, H., 1996 Interactive Exploratory Data Analysis Using MST-based Nonlinear Maping in: Faulbaum, F. & Bandilla, W. (Hrsg.) SoftStat ‘95 Lucius & Lucius Stuttgart 1996 1. Auflage Seite 123 - 128

Shepard, R. N., 1972 Introduction to Volume I in: Shepard, R. N., Romney, A. K., Nerlove, S. G. (Hrsg.) Multidimensional Scaling: Theory and Applications in the Behavioural Sciences, Volume I, Seminar Press New York 1972 1. Auflage Seite 1 - 22

Smith, P.W.F., 1992 Assessing the Power of Model Selection Procedures Used When Graphical Modelling in: Dodge, Y. & Whittaker, J. (Hrsg.) Coputational Statistics Volume I Physica Heidelberg 1992 1. Auflage Seite 275 - 280

Spangenberg, N. & Wolff, K.E., 1991 Comparison of Biplot Analysis and Formal Concept Analysis in the Case of a Repertory Grid in: Bock, H.H. & Ihm, P. (Hrsg.) Classification, Data Analysis, and Knowledge Organisation Springer Heidelberg 1991 1. Auflage Seite 104 - 112

Theus, M., 1996 Trellis Displays vs. Interactive Graphics in: Faulbaum, F. & Bandilla, W. (Hrsg.) SoftStat ‘95 Lucius & Lucius Stuttgart 1996 1. Auflage Seite 129 - 138

Thomas, E., 1992 Sortimentsanalyse bei Süßkirschen mit Hilfe mehrdimensionaler statistischer Verfahren in: Enke, H., Gölles, J., Haux, R., Wernecke, K.-D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Verlag Stutgart 1992 1. Auflage Seite 297 - 314

Tukey, P.A. & Tukey, J.W., 1981 Part III: Graphical Display of Data Sets in 3 or More Dimensions in: Barnett, V. (Hrsg.) Interpreting Multivariate Data Wiley New York 1981 1. Auflage Seite 189 - 278

Wille, R., 1982 Restructuring Lattice Theory: An Approach Based on Hierarchy of Concepts in: Rival, I. (Hrsg.) Ordered Sets Reidel Dordrecht-Boston 1982 1. Auflage Seite 445 - 470

Wolf, P., 1992 Zehn Punkte zur Standortbestimmung von EDA - Einige Bemerkungen über die Verbindlichkeit von exploratorischer Datenanalyse in: Enke, H., Gölles, J., Haux, R., Wernecke, K.-D. (Hrsg.) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften Gustav Fischer Verlag Stutgart 1992 1. Auflage Seite 317 - 328

Wolff, K. E., 1993 A First Course in Formal Concept Analysis in: Faulbaum, P. (Hrsg.) Softstat ‘93: Advances in Statistical Software 4 Lucius & Lucius Stuttgart 1993 1. Auflage Seite 429 - 438

Becker, R.A., Cleveland, W.S., Shuy, M-J., 1994 Trellis Display: A Framework for Visualising 2D and 3D Data AT&T Bell Laboratories, Statistics Research Report 8 Murray Hill 1994

Benne, R., 1990 Ergebnisse der Anwendung mehrdimensionaler Varianz- und Diskriminanzanalyse bei Versuchen zu Gemüse Archiv Gartenbau 1990 *38* Seite 49 - 65

Bitsch, V., 1994 Erfolgsanalyse bei Gartenbaubetrieben auf der Basis von Jahresabschlußdaten Forschungsberichte zur Ökonomie im Gartenbau Nr. 77 Fachbereich Gartenbau der Universität Hannover, Dissertation, Hannover und Weihenstephan 1994

Blecken, H. Betriebsbegleitende Untersuchungen bei Fuchsien 1987 Gartenbauberatungsring e.V. Hannover, Eigendruck Hannover 1987

BMDP, 1995 BMDP/Diamond User's Guide BMDP Inc. Los Angeles 1995

Bokelmann, W., 1987 Theoretischer Bezugsrahmen und empirische Untersuchungen zu Entscheidungsabläufen in der gärtnerischen Produktion Forschungsberichte zur Ökonomie im Gartenbau Nr. 63 Fachbereich Gartenbau der Universität Hannover, Dissertation, Hannover und Weihenstephan 1987

Bokelmann, W., 1993 Früherkennung von Unternehmenskrisen im Gartenbau auf der Grundlage von Jahresabschlußdaten Fachbereich Gartenbau der Universität Hannover, Habilitation, Hannover 1993

Bokelmann, W., & Voth, M., 1986 Poinsettien unter der Lupe Institut für Gartenbauökonomie der Universität Hannover, Eigendruck Hannover 1986

Borg, I. Groenen, P.J.F., 1997 Multitrait-Multimethod by Multidimensional Scaling Vortrag am 3.3.1997 auf der SoftStat '97 in Heidelberg

ESZ (Ernst Schröder Zentrum für begriffliche Wissensverarbeitung e.V.), 1996 Einführung in die begriffliche Wissensverarbeitung; Seminarunterlage vom 10.10. und 11.10.1996, TH Darmstadt

Görgens, M., 1991 Die Anwendung betriebsbegleitender Untersuchungen im Obstbau Arbeitsbericht Nr. 67 Institut für Gartenbauökonomie der Universität Hannover, Hannover 1991

Gottschlich, W. Empirische Identifikation von typischen Schwachstellenprofilen landwirtschaftlicher Unternehmen FAA, Bonn 1995

Gower, J.C., 1997 Email von J.C. Gower an Stefan Krusche vom 30.12.1997

Inselberg, A., 1997 Parallel Coordinates: Multidimensional Visualisation and Multivariate Applications Vortrag am 5.3.1997 auf der SoftStat '97 in Heidelberg 1997

Jokiel, A., 1996 Persönliche Mitteilung von Zierpflanzenbauberater Alfons Jokiel der Bezirksstelle für Gartenbau, Münster, der Landwirtschaftskammer Westfalen-Lippe April, 1996

Karaman, Z., 1995 Procedure DPARALLEL Genstat 5 Procedure Library Manual Release 3[2] 1995

Kollewe, W., Skorsky, M., Vogt, F., Wille, R., 1994 TOSCANA - ein Werkzeug zur begrifflichen Analyse und Erkundung von Daten Technische Hochschule, Darmstadt Preprint-Nr. 1636 1994

Kreiner, S., 1989 Graphical Modelling Using DIGRAM Forschungsbericht 89/11 Universität Kopenhagen 1989

Kühne, H., 1997 Persönliche Mitteilung von Betriebswirtschaftsberater Hans Kühne des Referat Gartenbau, Münster, der Landwirtschftskammer Westfalen-Lippe, Februar, 1997

NAVICON, 1996 TOSCANA 2.1 Benutzerhandbuch Navicon Hanau 1996

Ollerton, J. & Harding, S.A., 1995 Procedure BANK Genstat 5 Procedure Library Manual Release 3[2] 1995

Peters, M. Cyclamen-Praxisversuch - Ergebnisse des Haltbarkeitsversuchs Landwirtschaftskammer Westfalen-Lippe, Eigendruck Münster 1994

Uhte, R., 1997 Bereitstellung von Daten Schreiben von Herrn Dr. Uhte vom 7.4.1997

Wiesmann, R., 1985 Beurteilung des Beratungseffekts betriebsbegleitender Untersuchungen in Papenburger Gemüsebaubetrieben Fachgebiet Betriebslehre des Gartenbaus der Universität, Hannover, Diplomarbeit 1985

Wille, R., 1987 Bedeutungen von Begriffsverbänden Technische Hochschule, Darmstadt Preprint-Nr. 1058 1987

Wolff, K. E., 1988 Einführung in die formale Begriffsanalyse Forschungsbericht der Forschungsgruppe Begriffsanalyse Darmstadt 1988

Wollf, K. E. & Stellwangen, M., 1993 Conceptual Optimization in the Production of Chips Forschungsbericht der Forschungsgruppe Begriffsanalyse Darmstadt 1993

Verzeichnis der Abkürzungen und Symbole

zum ersten Mal verwendet auf Seite ...	Abkürzung oder Symbol	Bedeutung
4	EDA	Exploratory Data Analysis
4	IDA	Initial Data Analysis
4	CDA	Confirmtory Data Analysis
8	n	Anzahl Objekte (i = 1 ... n)
8	p	Anzahl Variablen (j = 1 ... p)
9	$\mathbf{X}$	(n x p) Datenmatrix
9	$\mathbf{S}$	(p x p) Kovarianzmatrix
9	$\mathbf{R}$	(p x p) Korrelationsmatrx
9	q	Anzahl Dimensionen in der Dimensionserniedrigung (q < p, j = 1 ... q))
10	f_q	Velicers f-Wert bei q Komponenten
10	m	Anzahl an Dimensionen in der Kreuzvalidierung
10	$\hat{\beta}^{(m)}$	allgemeiner Parameter bei m Dimensionen in der Kreuzvalidierung
10	$\mathbf{x_i}$	(1 x p) (Zeilen)Vektor der p Variablenwerte von Objekt i
10	$\hat{\mathbf{x}}_\mathbf{i}$	geschätzter (1 x p) Vektor der p Variablenwerte von Objekt i bei Betrachtung von m Dimensionen in der Kreuzvalidierung
10	E(m)	Diskrepanzmaß bei Betrachtung von m Dimensionen im Vergleich zu p Dimensionen in der Kreuzvalidierung
10	k	allgemeiner Korrekturfaktor in der Kreuzvalidierung
11	$\mathrm{PRESS}^{(m)}$	Prediction Sum of Squares bei Betrachtung von m Dimensionen in der Kreuzvalidierung
11	x_{ij}	Wert von Objekt i bei Variable j
11	$\hat{x}_{ij}^{(m)}$	geschätzter Wert von Objekt i bei Variable j bei Betrachtung von m Dimensionen in der Kreuzvalidierung

11	Q_α	kritischer Wert für Residuen in der Hauptkomponentenanalyse
11	l_j	Eigenwert von Hauptkomponente j
11	c_α	Wert der Standardnormalverteilung bei Irrtums-wahrscheinlichkeit von α
17	s_{rt}	Ähnlichkeit zwischen zwei Objekten r und t
17	d_{rt}	Unähnlichkeit zwischen zwei Objekten r und t
17	x_{rj}, x_{tj}	Wert der Variablen j bei Objekt r beziehungsweise t
17	u_{rtj}, v_{rtj}	aus Datenmatrix abgeleitete Werte zur Berechnung des allgemeinen Ähnlichkeitskoeffizienten für die Objekte r und t bei Variable j
23	c_{rt}	Anzahl der Übereinstimmungen bei r und t, geteilt durch p
23	a, b, c, d	Anzahl Fälle in der Zwei-Wege-Tafel zur Berechnung von Proximitätsmaßen für binäre Variablen
17	**D**	(n x n) Proximitätsmatrix
17	G	Verhältnis von Eigenwerten
18	**B**	(n x n) Matrix $\mathbf{XX'}$
18	r, t, u	drei Objekte
18	q	Anzahl an Dimensionen in der Dimensionserniedrigung (siehe Seite 9, in der mehrdimensionalen Skalierung aber q < n)
19	δ_{rt}	Dissimilaritäten (Unähnlichkeiten der Ausgangs-proximitätsmatrix)
19	d_{rt}	Distanzen (die euklidischen Distanzen der Objekte in der durch die mehrdimensionale Skalierung erzielten Konfiguration in q Dimensionen)
20	$\hat{d}_{rt}$	Disparität (der Schätzwert, der durch die dem mehrdimensionalen Skalierungsmodell zugrunde gelegte Beziehung von δ_{rt} und d_{rt} geschätzt wird)
20	q_{max}	maximale Anzahl an Dimensionen, die in der mehrdimensionalen Skalierung betrachtet wird beziehungsweise in der Faktoranalyse betrachtet werden kann (Seite 27)
24	**Z**	(k x p) Kontingenztafel

24	k	Anzahl Zeilen von **Z** (i = 1 ... k)
24	p	Anzahl Spalten von **Z** (j = 1 ... p)
24	**z**	(k x 1) (Spalten)Vektor der Zeilensummen von **Z**
24	**s**	(1 x p) (Zeilen)Vektor der Spaltensummen von **Z**
24	$\mathbf{D}_z$	Diagonalmatrix der Zeilensummen mit k Zeilen
24	$\mathbf{D}_s$	Diagonalmatrix der Spaltensummen mit p Zeilen
24	$\mathbf{Z}^w$	mit $\mathbf{D}_z$ und $\mathbf{D}_s$ gewichtete Matrix **Z**
24	**U**	(k x p) Matrix der linken singulären Vektoren der Eigenwertzerlegung von $\mathbf{Z}^w$
24	**V**	(p x p) Matrix der rechten singulären Vektoren der Eigenwertzerlegung von $\mathbf{Z}^w$
24	$\mathbf{D}_\alpha$	Diagonalmatrix der singulären Werte der Eigenwertzerlegung von $\mathbf{Z}^w$ mit p Zeilen
24	N	Summe aller Werte in **Z**
24	$\mathbf{D}_r$	Diagonalmatrix der Zeilensummen von (1/N)**Z** mit k Zeilen
24	$\mathbf{D}_c$	Diagonalmatrix der Spaltensummen von (1/N)**Z** mit p Zeilen
24	**F**	Koordinaten der Zeilenprofile
24	**G**	Koordinaten der Spaltenprofile
25	**Φ**	standardisierte Koordinaten der Zeilenprofile
25	**Γ**	standardisierte Koordinaten der Spaltenprofile
25	**I**	(p x p) oder (k x k) Einheitsmatrix
26	$\mathbf{Z}^I$	Indikatormatrix
27	$\mathbf{Z}^B$	Burt-Matrix
27	α	Inertia (singulärer Wert) in der Korrespondenzanalyse
28	z_{ij}	Boniturwert von Objekt i (i = 1 ... n) bei Variable p (1 ... j) in ‘verdoppelter’ Matrix
28	t_j	obere Grenze der Bonitursкala für Variable j
28	j+	Plusspalte der ‘verdoppelten’ Matrix der Korrespondenzanalyse für Variable j

28	j-	Minusspalte der ‘verdoppelten’ Matrix der Korrespondenzanalyse für Variable j
28	$\tilde{k}_j$	mit oberer Grenze der Bonitursкala gewichteter mittlerer Boniturwert von Variable j
28	$\bar{z}_j$	mittlerer Boniturwert von Variable j
28	pol_{mj}	Polarisation des Mittels
28	k_{ij}	Boniturwert z_{ij} gewichtet mit t_j
28	pol_{ob}	Polarisation der Objekte
28	S_j	Standardabweichung von Variable j
29	fac_j	Transformationsfaktor für Variable j
29	z_{ij}^*	mit fac_j transformierter Wert für z_{ij}
31	x_i, x_i^*, y_j	drei Variablen
31	$r_{x_i x_i^*}$	Korrelation zwischen x_i und x_i^*
31	$r_{x_i x_i^* \bullet y_j}$	partielle Korrelation zwischen x_i und x_i^*, gegeben y_j
34	**Y**	(n x p) Matrix der Hauptkomponentenwerte von **X**
34	d_{rt}^p	euklidische Distanz zwischen zwei Objekten r und t in p Dimensionen
34	d_{rt}^q	euklidische Distanz zwischen zwei Objekten r und t in q Dimensionen
34	r_i	Differenz zwischen d_{rt}^p und d_{rt}^q
34	R	p-dimensionaler Raum von X
34	L	q-dimensionaler Unterraum von R
35	$\bar{x}_j$	Mittelwert von Variable j
35	val	beliebiger Wert
35	*i*	$i = \bar{x}_j + val$
35	$\mathbf{m}_{\mathbf{int}}^{ij}$	(1 x q) Zeilenvektor der Koordinaten für Interpolationsmarker für Variable j bei Wert *i*
35	$\mathbf{m}_{\mathbf{pred}}^{ij}$	(1 x q) Zeilenvektor der Koordinaten für Prediktionsmarker für Variable j bei Wert *i*

35	$\mathbf{e_j}$	(1 x p) Zeilenvektor mit einer 1 bei Variable j und sonst nur Nullen
35	$\mathbf{A_q}$	(p x q) Matrix der Eigenvektoren nach Dimensionserniedrigung von **X**
35	***v***	Index für Markerabstände
35	$\mathbf{m}_{\mathbf{int}}^{vj}$	(1 x q) Zeilenvektor der Koordinaten für Interpolationsmarker für Variable j bei Wert ***v***
35	$\mathbf{m}_{\mathbf{pred}}^{vj}$	(1 x q) Zeilenvektor der Koordinaten für Prediktionsmarker für Variable j bei Wert ***v***
36	a_{jj^*}	Elemente der (p x q) Matrix der Eigenvektoren $\mathbf{A_q}$
36	con_{jj^*}	Beitrag der Variable j zum Eigenwert der Hauptkomponente j^*
37	$\mathbf{X}^*$	aus Proximitätsmatrix **D** (Seite 17) berechnete Koordinatenmatrix
37	**Q**	(p x p) Transformationsmatrix
38	CLP	Category Level Point (Kategorien-Stufen-Punkte)
39	***O***	Schnittpunkt der Biplotbahnen
39	**L**	(n x n) Matrix der Eigenwerte von $\mathbf{B} = \mathbf{X}^*\mathbf{X}^{*'}$
39	**1**	(1 x n) Zeilenvektor mit Einsen (oder (n x 1) siehe Fußnote 23))
39	$\mathbf{d}_{\mathbf{n+1}}^{v}$	(1 x n) Zeilenvektor der quadrierten Distanzen eines Pseudoobjekts zu den übrigen Objekten
42	CMP	Column Preserving
42	RMP	Row Preserving
45	A, B	zwei Gruppen, die mit der Hauptkomponentenanalyse analysiert werden
45	l_{jA}	Eigenvektoren von Gruppe A ($j_A = 1 \ldots q_A$) ($q_A \leq p$)
45	m_{jB}	Eigenvektoren von Gruppe B ($j_B = 1 \ldots q_B$) ($q_B \leq p$)
46	**L**	(q_A x p) Matrix der Eigenvektoren von Gruppe A
46	**M**	(q_B x p) Matrix der Eigenvektoren von Gruppe B
46	**N**	(q_A x q_A) Matrix **LM'ML'**

46	α_1	der kleinste Winkel zwischen einem beliebigen Vektor der ersten q Hauptkomponenten von A und dem am parallesten gelegenen Vektor der ersten q Hauptkomponenten von B
47	$\mathbf{a_j}$	Eigenvektor zum Eigenwert l_j von **N**
47	$\mathbf{b_j}$	ist gleich $\mathbf{L'a_j}$
47	$\mathbf{c_1 \ldots c_q}$	mittlere Komponenten der Dimensionen von A und B
47	$\mathbf{L_t}$	(q x p) Matrix mit den Eigenvektoren der Hauptkomponentenanalyse der Gruppe t (t = 1 ... g)
47	**H**	(p x p) Matrix $\sum_{t=1}^{g} \mathbf{L_t'L_t}$
48	δ_{tj}	Maß für die Abweichung der Gruppen t von den Vektoren, die einen Unterraum des Ausgangsdatenraums beschreiben, der allen Gruppen gleichzeitig so nahe wie möglich ist (mit t = 1 ... g und j = 1 ... q ($q \leq p$)).
48	$\boldsymbol{\xi_j}$	A priori festgegter Eigenvektor j (j = 1 ... p)
48	$\mathbf{b_j}$	typischer Eigenvektor j (j = 1 ... p)
48	$\mathbf{a_{tj}}$	Eigenvektor zum Eigenwert l_j von **H** (j = 1 ... p) für Gruppe t (t = 1 ... g)
50	$\mathbf{X}$, $\mathbf{X^*}$	zwei (n x p) Matrizen
50	x_{ij}, x_{ij}^*	Elemente von **X** und $\mathbf{X^*}$ für Objekt i und Variable j
50	M^2	Summe der quadrierten Abweichungen von x_{ij} und x_{ij}^*
50	$\mathbf{G_X}$, $\mathbf{G_{X^*}}$	Zentroid (Mittelwertsvektor) von **X** beziehungsweise $\mathbf{X^*}$
50	c	Dilationsfaktor
51	$\mathbf{X_t}$	(n x p_t) Matrix für t = 1... g mit j = 1 ... p
51	A_{it}	von Objekt i in Gruppe t besetzter Punkt im Koordinatensystem von $\mathbf{X_t}$
51	$x_{i1t}, x_{i2t}, \ldots, x_{ipt}$	Koordinaten von $\mathbf{X_t}$
51	G	Zentroid aller Punkte allerGruppen

51	G_t	Zentroid von Gruppe t
51	F_i	Zentroid von Objekt i für alle Konfigurationen t
51	*O*	gemeinsamer Ursprung aller Gruppen nach Translation
51	g	Anzahl Gruppen
52	w_{tj}	Gewichtungswerte je Gruppe t und Dimension j
52	O_t	Maß der Anpassungsgüte von Konfiguration für Gruppe t und Gesamtkonfiguration
53	SSP	Sum of Squares and Products
53	**B**	(p x p) SSP-Matrix zwischen den Gruppen
53	**W**	(p x p) SSP-Matrix innerhalb der Gruppen
56	g	Gegenstände
56	G	Menge der Gegenstände g
56	m	Merkmale
56	M	Menge der Merkmale
56	K	ein formaler Kontext
56	I	binäre Relation zwischen den Elementen der Mengen G und M
56	A	ein Begriffsumfang
56	B	ein Begriffsinhalt
56	$\subseteq$	Teilmenge von ...
59	b_1, b_2	Unterbegriff, Oberbegriff
63	A, B, C	drei Variablen
63	$\perp\!\!\!\perp$	... unabhängig ...
63	|	... gegeben ...
63	EH-	Edwards-Havránek
64	*I*	diskrete Zufallsvariable
64	*i*	Wert von *I*
64	p_i	Wahrscheinlichkeit, daß *I* den Wert *i* annimmt
64	Y	kontinuierliche Zufallsvariable
64	μ_i	Mittelwert von Y, gegeben *i*
64	$\sum_i$	Kovarianzmatrix von Y, gegeben *i*

65	n_{jkl}	beobachtete Zellhäufigkeit einer Drei-Wege Tafel der diskreten Variablen A, B und C mit den Klassen j = 1 ... a, k = 1 ... b, und l = 1 ... c
65	$\hat{m}^0_{jkl}$	Maximum-Likelihood Schätzer der Zellhäufigkeit des komplexeren Modells
65	$\hat{m}^1_{jkl}$	Maximum-Likelihood Schätzer der Zellhäufigkeit des einfacheren Modells
65	ln	natürlicher Logarithmus
65	G^2	$G^2 = 2\sum_{jkl} n_{jkl} \ln\left(\hat{m}^1_{jkl} / \hat{m}^0_{jkl}\right)$
67	CART	Classification And Regression Trees
68	CHAID	Chi-Square Automatic Interaction Detector
70	$f_{x_{ij}}(\omega)$	Andrews Funktion
70	π	3,1415927
70	ω	$-\pi \le \omega \le \pi$
73	AWE	Approximate Weight of Evidence
73	s(i)	Silhouettenbreite
73	a(i)	mittlere Unähnlichkeit von Objekt i zu dem Cluster, dem es zugeordnet ist
73	b(i)	kleinste Unähnlichkeit von Objekt i zu allen Clustern
73	d(i)	mittlerer Unähnlichkeit von Objekt i zu dem Cluster mit dem es zuerst verschmolzen wird
74	AC	Agglomerative Coefficient
76	a, b	obere und untere Grenzen einer Gleichverteilung
76	ϑ	Zufallszahlen der Gleichverteilung mit den Grenzen a und b
81	W, W_R, W_L, W_G	Teststatistiken nach HARRIS, 1985
82	$\mathbf{S}$	Maximum-Likelihood Schätzer für die Kovarianzmatrix $\mathbf{\Sigma}$ von $\mathbf{X}$
82	w_i	Gewicht von Objekt i
82	$\mathbf{x_i}$	Wertevektor für Objekt i über alle p Variablen
82	$\bar{\mathbf{x}}$	Mittelwertsvektor von $\mathbf{X}$

82	$\bar{\mathbf{x}}_{\mathbf{M}}$	robuster Schätzer des Mittelwertsvektors von **X**
82	$\mathbf{S}_{\mathbf{M}}$	robuster Schätzer der Kovarianzmatrix von **X**
82	dm_i	Mahalanobis-Distanz für Objekt i vom Mittelwertsvektor von **X**
83	α, β	Parameter für die Schätzung der robusten Kovarianzmatrix
83	p	Anzahl der Variablen in **X**
83	dm_0	$dm_0 = \sqrt{p} + \alpha\sqrt{2}$
93	MTMM	Multi-Trait-Multi-Method
106	mca	Chi-Quadrat-Distanz
106	emc	Extended Matching Coefficient

Anhang Teil I A

Abbildungen zur Auswertung der betriebsbegleitenden Untersuchung bei Cyclamen, Kapitel 3.1

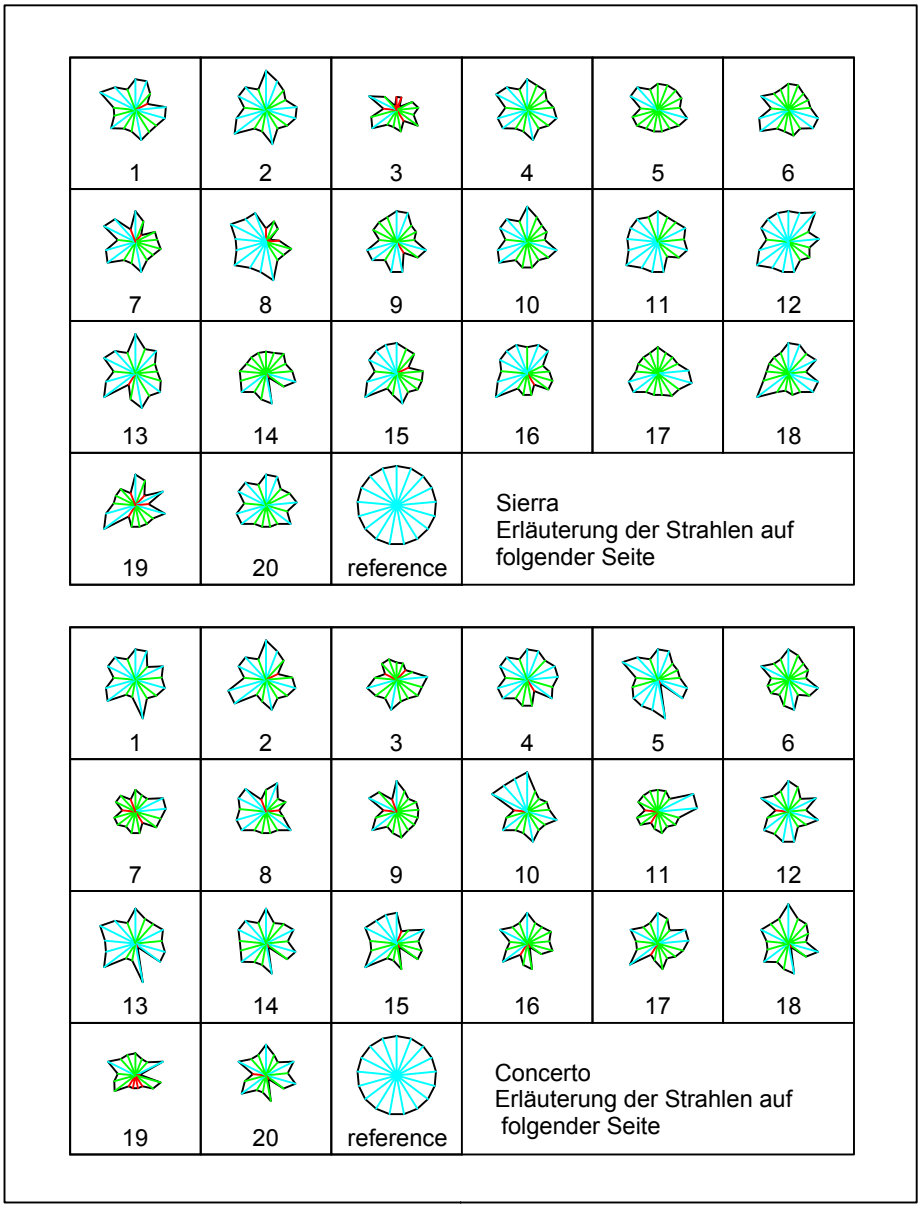
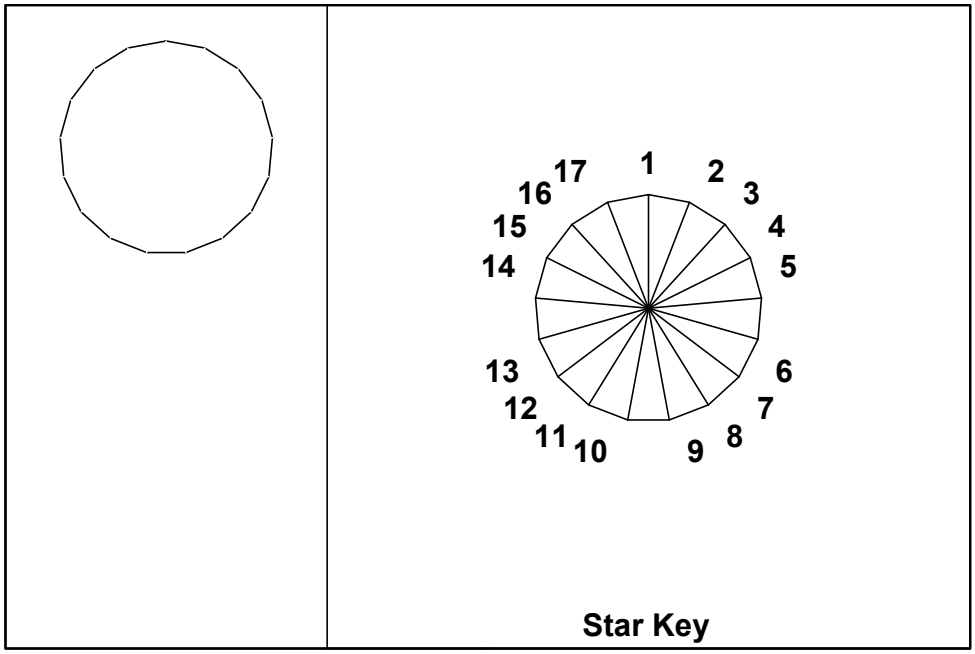


Abbildung A1: Starplots der Boniturwerte aller Qualitätsmerkmale aller Betriebe für ‘Sierra’ und ‘Concerto’



Erläuterung der Strahlen in den Starplots der Boniturwerte aller Qualitätsmerkmale aller Betriebe für ‘Sierra’ und ‘Concerto’

1 Gesamteindruck Woche 44
2 Gesamteindruck Woche 46
3 Gesamteindruck Woche 48
4 Wurzelbild Woche 44
5 Wurzelbild Woche 48
6 Knospenbesatz Woche 44
7 Knospenbesatz Woche 46
8 Knospenbesatz Woche 48
9 Welke Woche 44
10 Welke Woche 46
11 Welke Woche 48
12 Vergilbung Woche 44
13 Vergilbung Woche 46
14 Vergilbung Woche 48
15 Krankheiten Woche 44
16 Krankheiten Woche 46
17 Krankheiten Woche 48

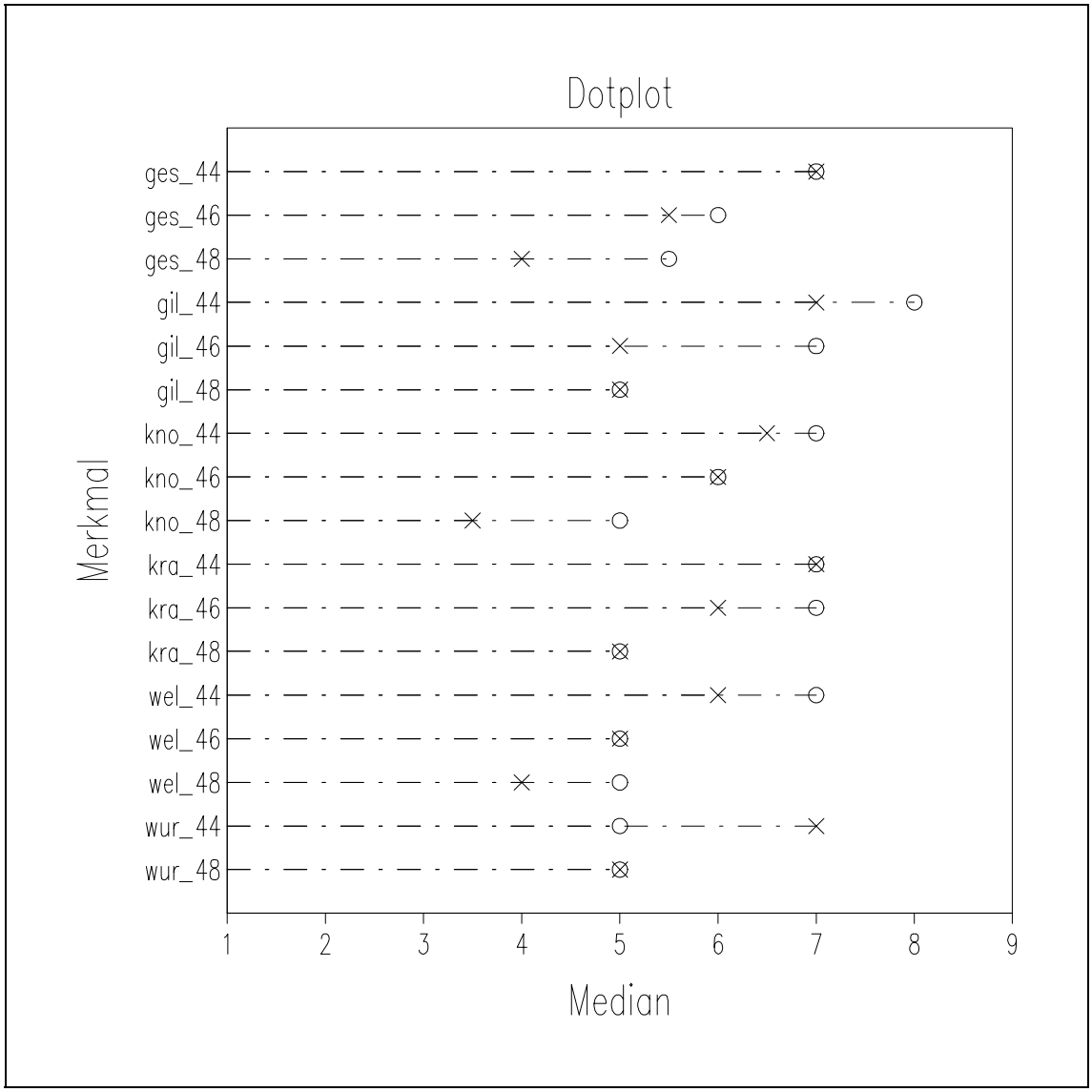


Abbildung A2: Dotplot der Mediane aller Qualitätsmerkmale (über alle Betriebe) für ‘Sierra’ (Kreis) und ‘Concerto’ (Kreuz)

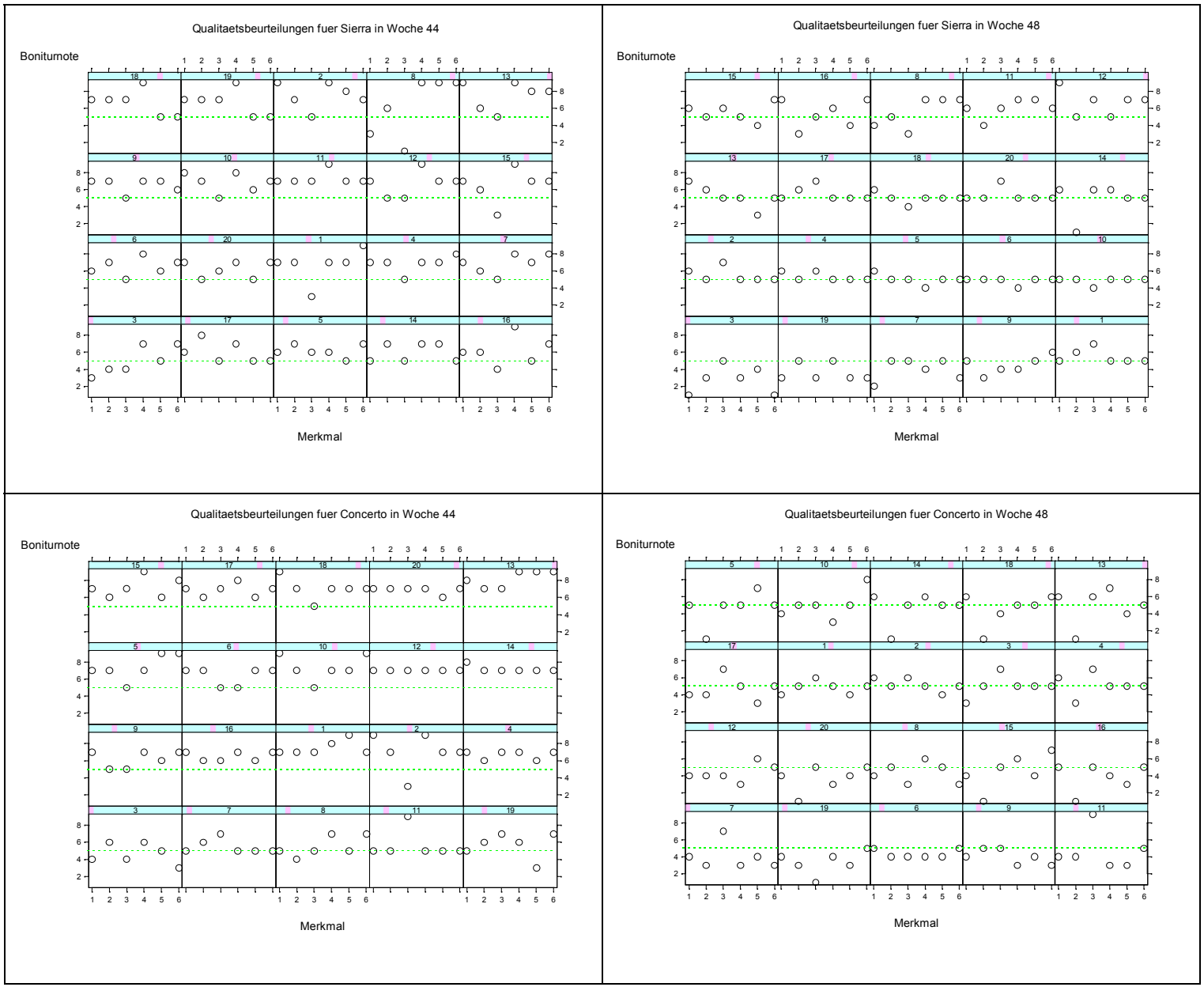


Abbildung A3: Trellis Display mit xy-Plot, alle Merkmale, konditioniert nach Betrieb, Woche 44 und Woche 48

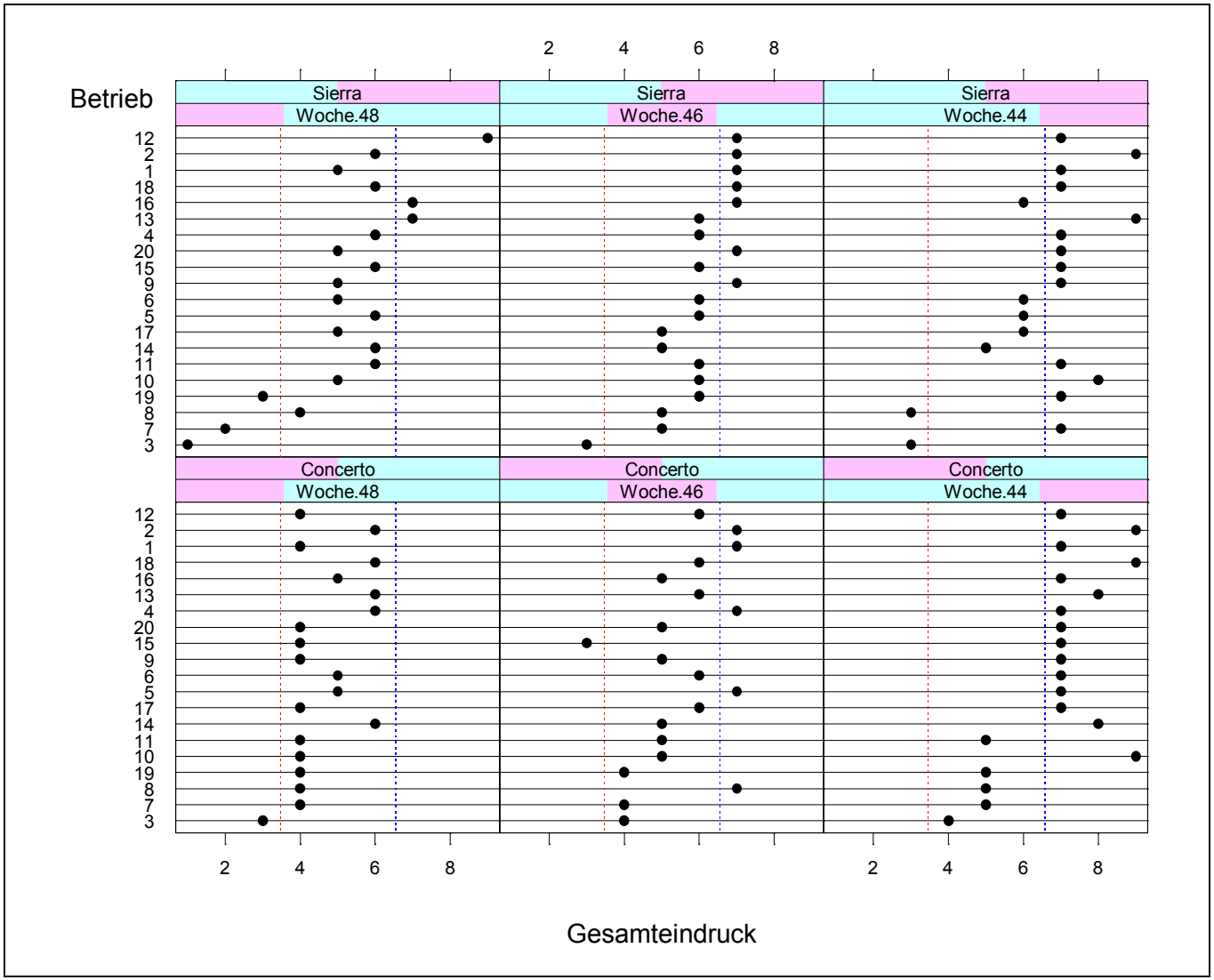


Abbildung A4: Trellis Display mit Dotplot, Beurteilung Gesamteindruck je Betrieb, konditioniert nach Woche und Sorte

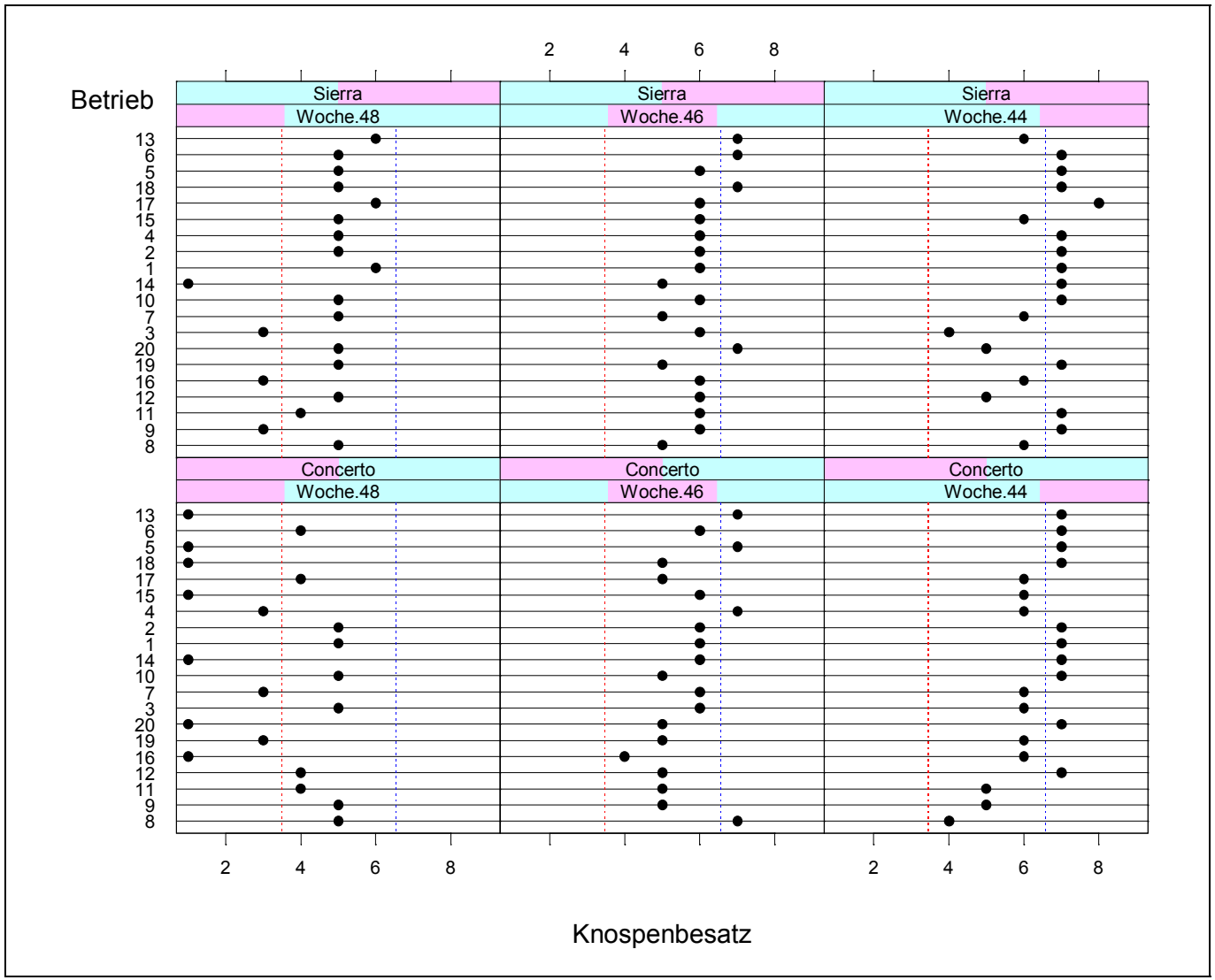


Abbildung A5: Trellis Display mit Dotplot, Beurteilung Knospenbesatz je Betrieb, konditioniert nach Woche und Sorte

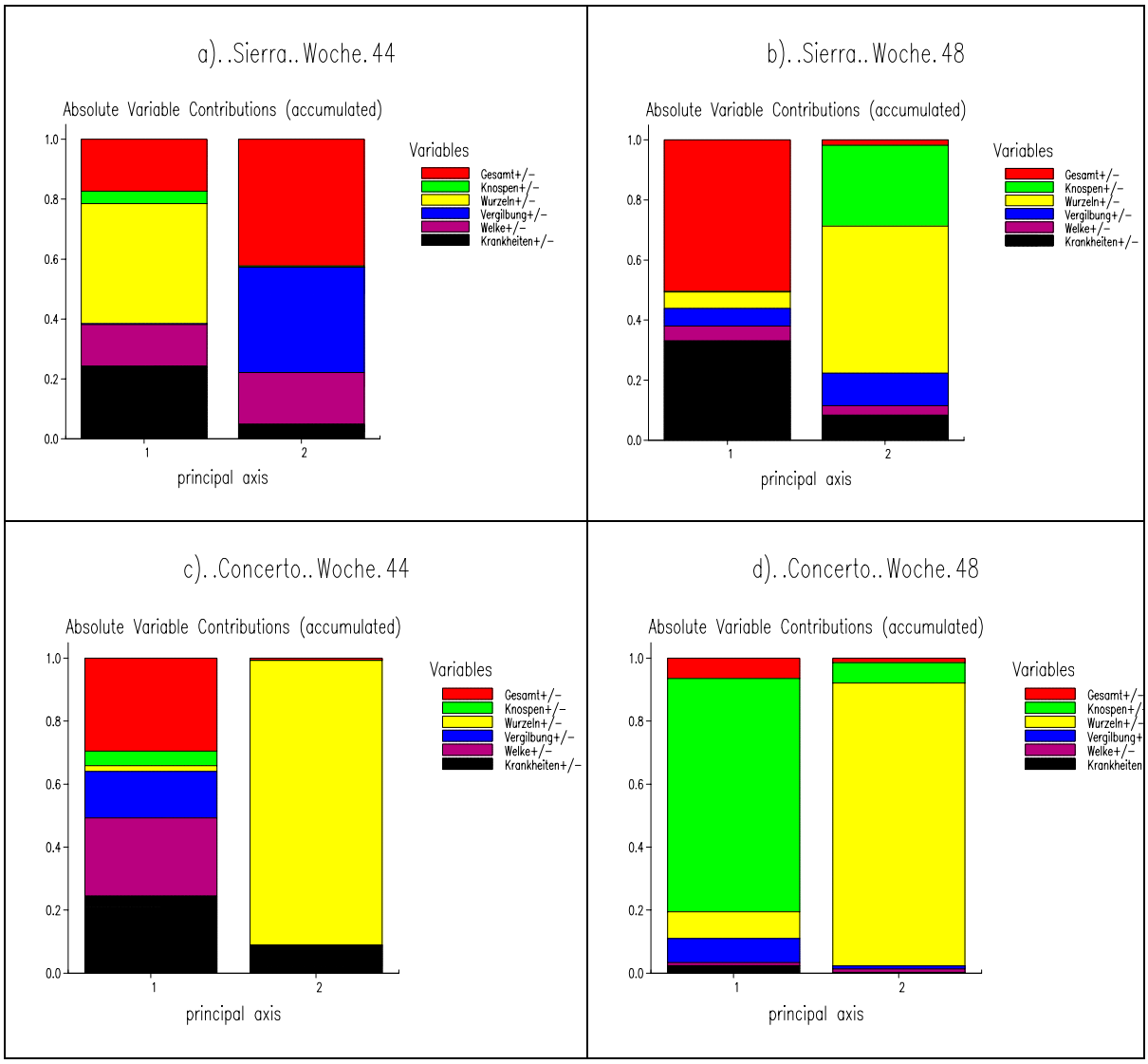


Abbildung A6 a,b,c,d: Kumulierte absolute Beiträge der Variablen; a) bei ‘Sierra’ in Woche 44, b) bei ‘Sierra’ in Woche 48, c) bei ‘Concerto’ in Woche 44, d) bei ‘Concerto’ in Woche 48

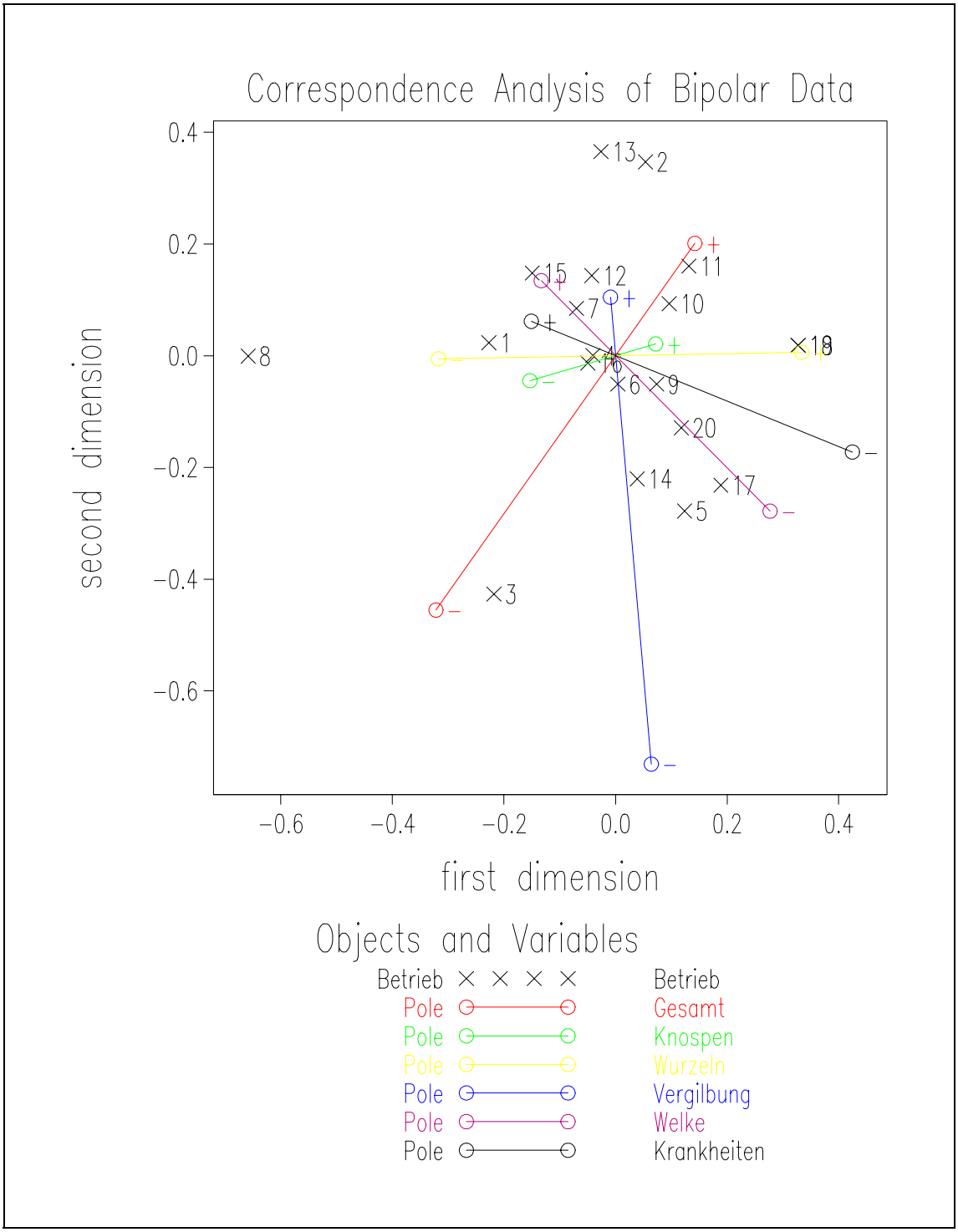


Abbildung A7: Korrespondenzanalyse bipolarer Daten, ‘Sierra’ Woche 44; Anteil der durch die erste Dimension erklärten Inertia 37,5%, Anteil der durch die zweite Dimension erklärten Inertia 30,6%

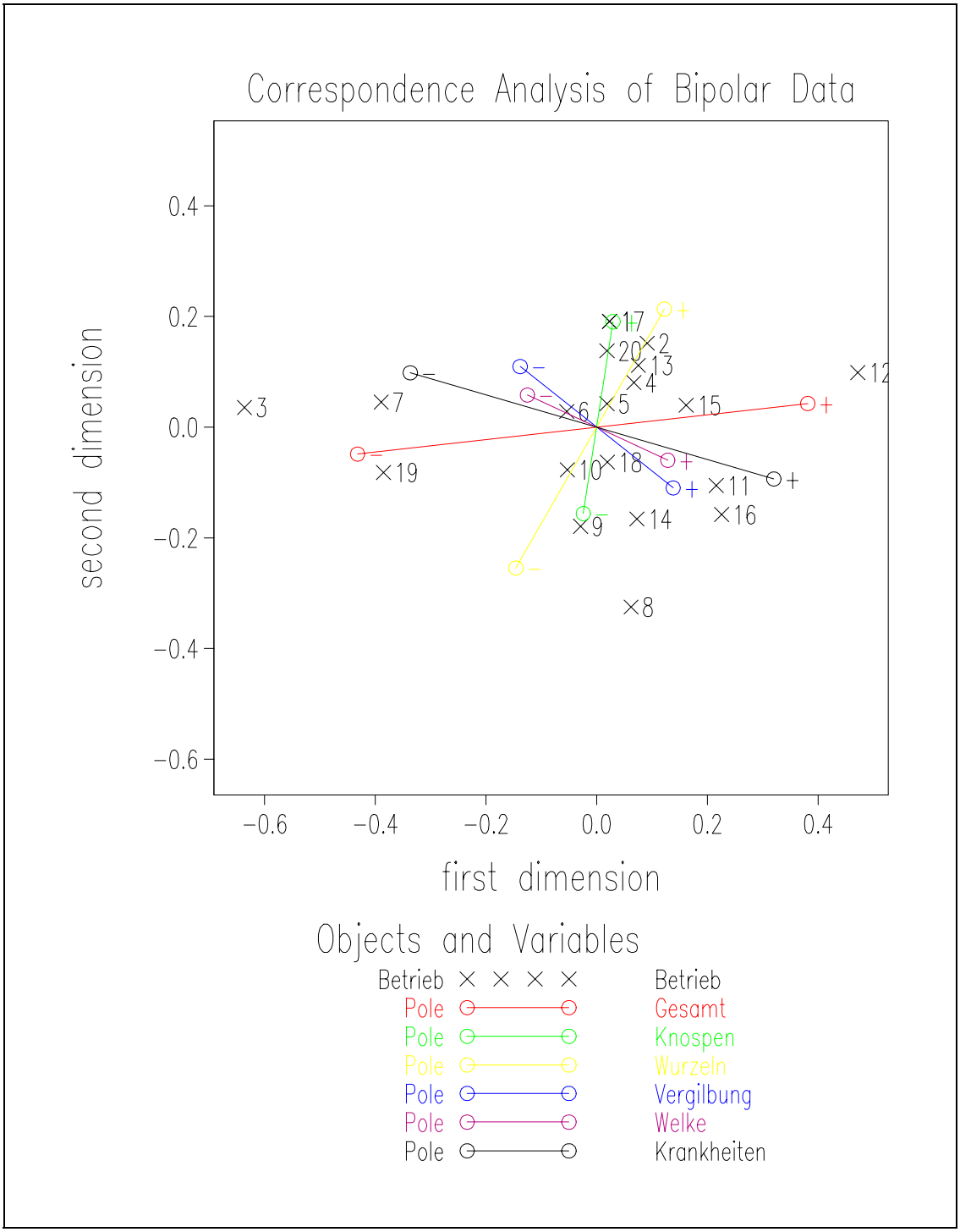


Abbildung A8: Korrespondenzanalyse bipolarer Daten, ‘Sierra’ Woche 48; Anteil der durch die erste Dimension erklärten Inertia 51,0%, Anteil der durch die zweite Dimension erklärten Inertia 17,3%

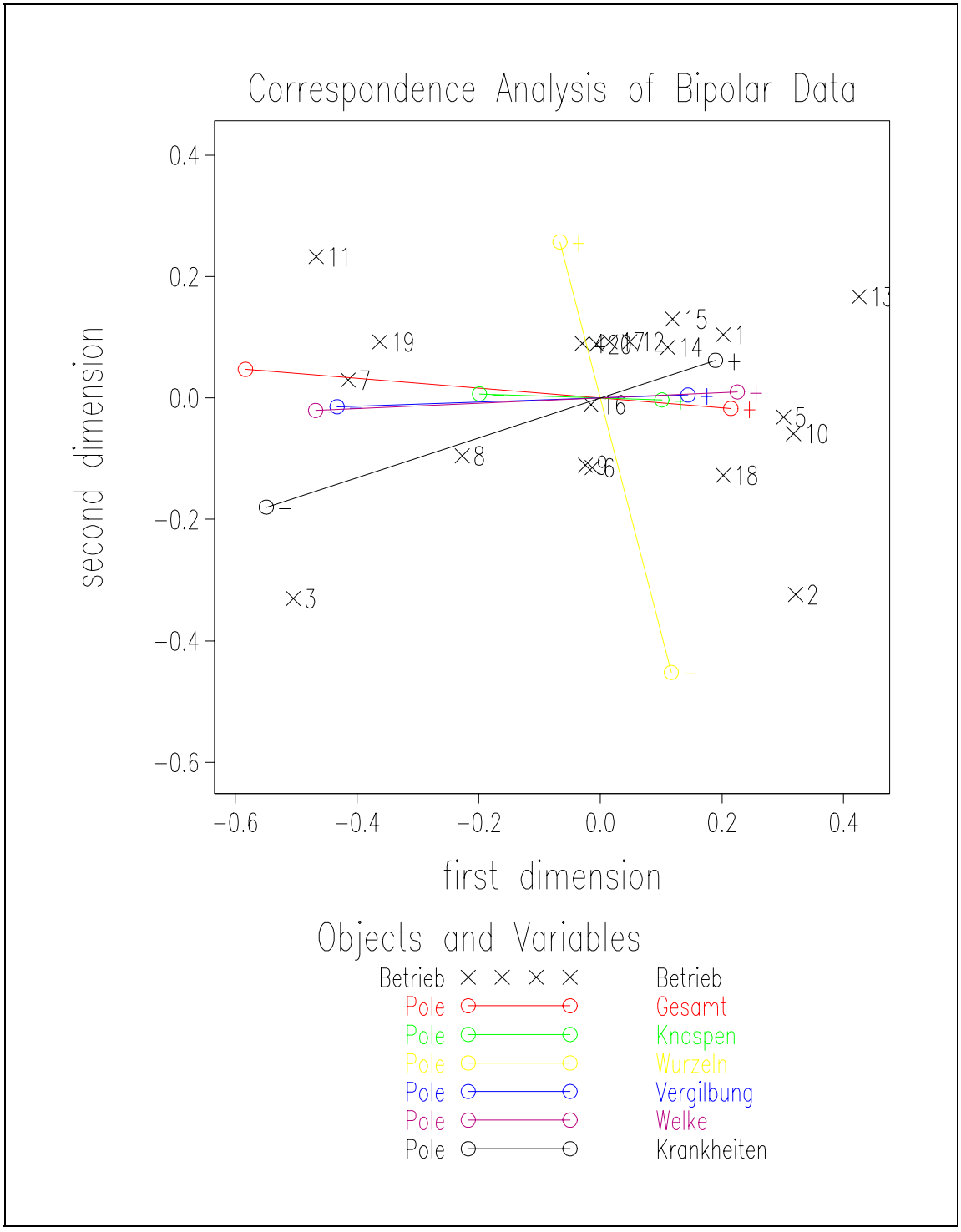


Abbildung A9: Korrespondenzanalyse bipolarer Daten, ‘Concerto’ Woche 44; Anteil der durch die erste Dimension erklärten Inertia 56,5%, Anteil der durch die zweite Dimension erklärten Inertia 17,1%

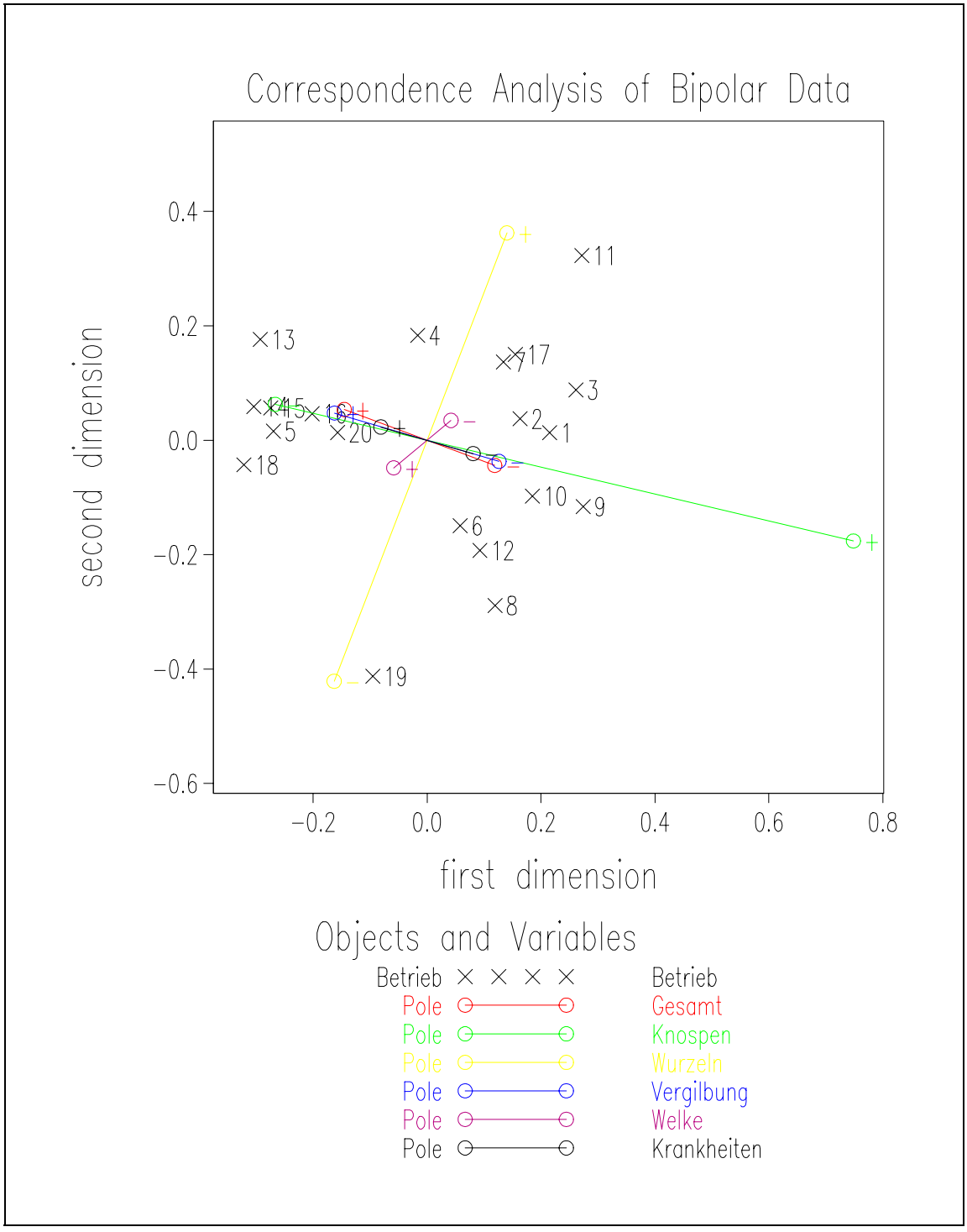


Abbildung A10: Korrespondenzanalyse bipolarer Daten, ‘Concerto’ Woche 48; Anteil der durch die erste Dimension erklärten Inertia 38,8%, Anteil der durch die zweite Dimension erklärten Inertia 24,5%

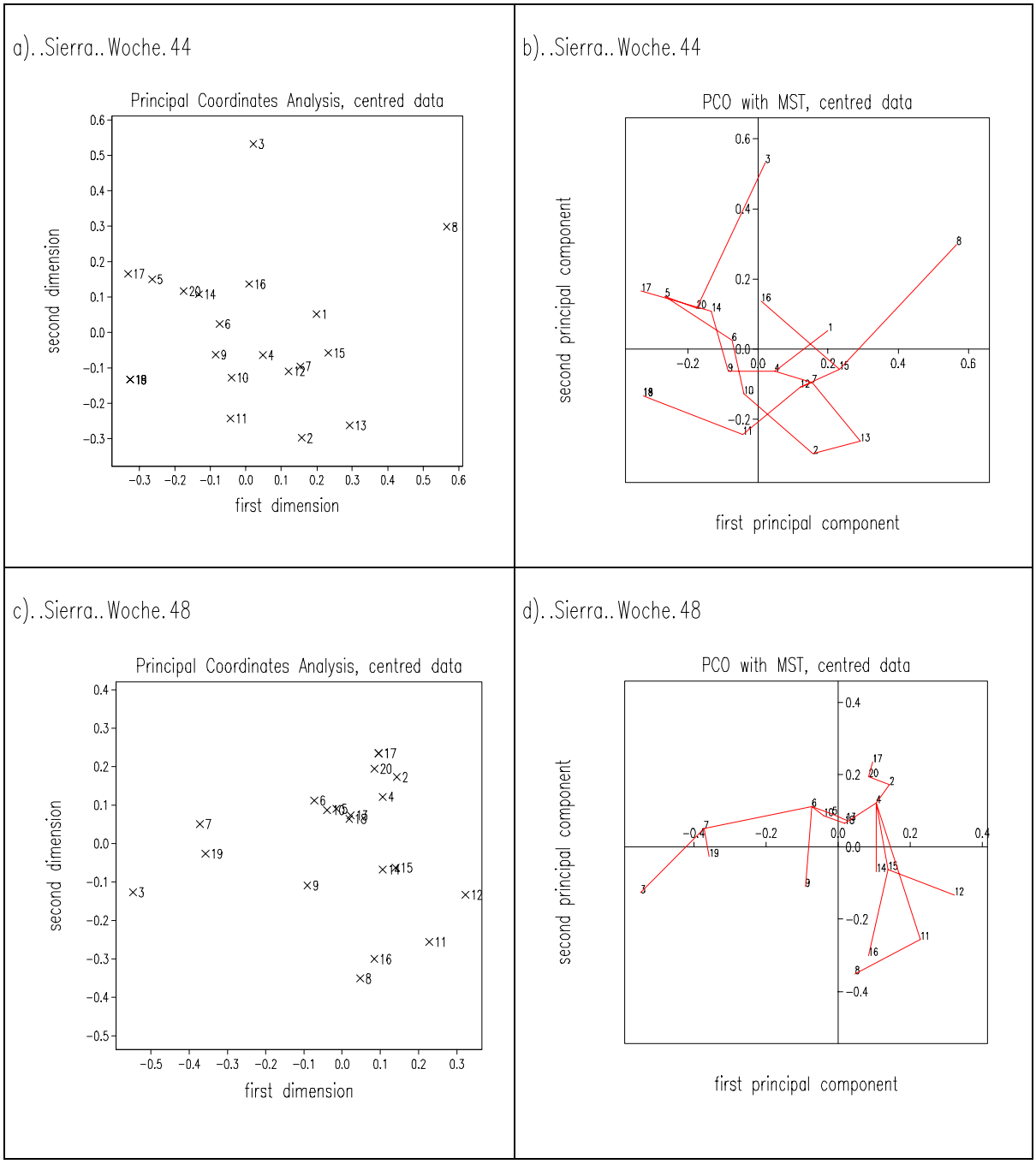


Abbildung A11 a,b,c,d: Hauptkoordinatenanalyse; Konfiguration in den ersten beiden Dimensionen, mit und ohne überlagerten Multiple Spanning Tree; a) und b) bei ‘Sierra’ in Woche 44 (erklärte Varianz in der ersten Dimension 24,6%, in der zweiten Dimension 18,5%); c) und d) bei ‘Sierra’ in Woche 48 (erklärte Varianz in der ersten Dimension 25,5%, in der zweiten Dimension 17,4%)

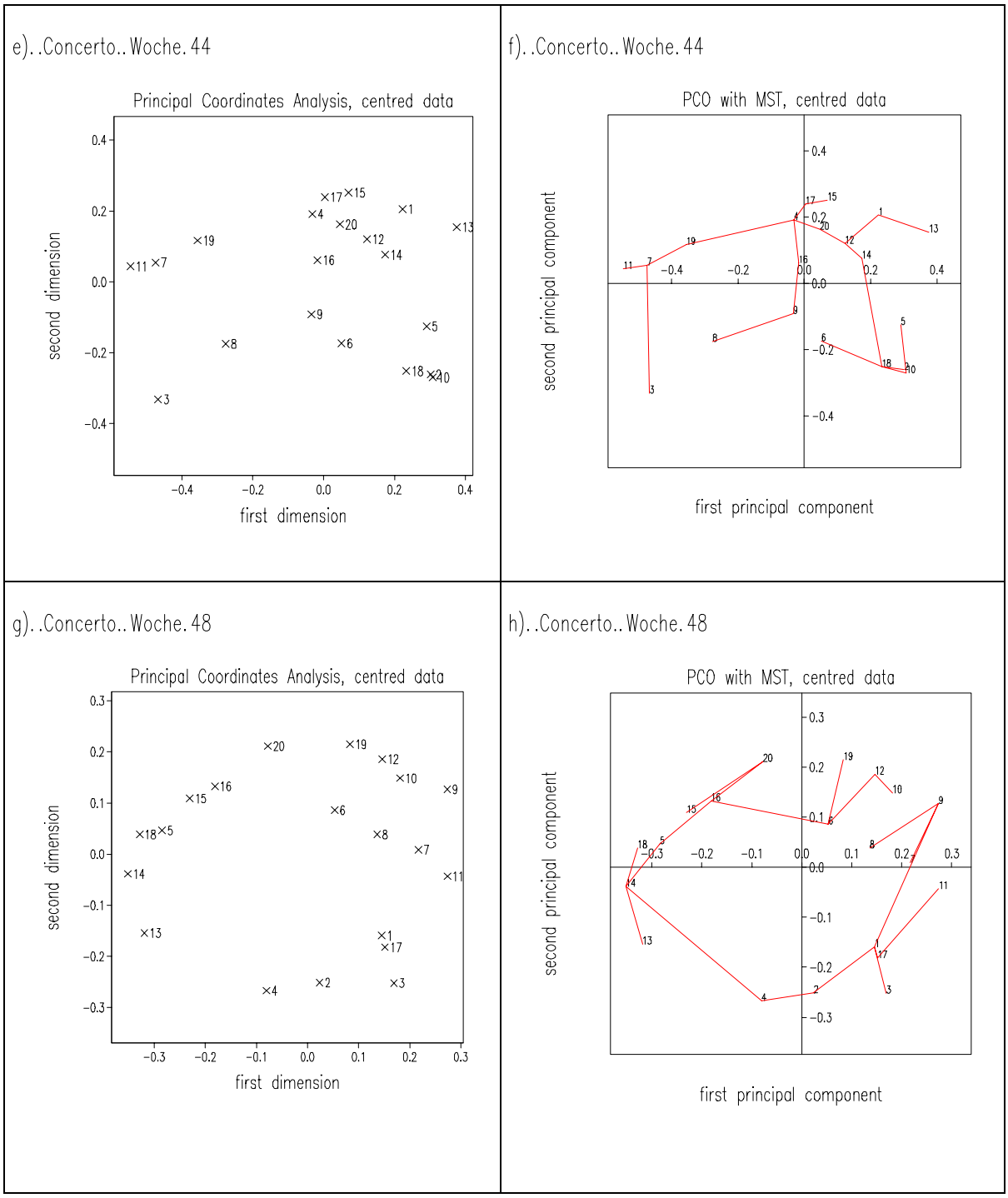


Abbildung A11 e,f,g,h: Hauptkoordinatenanalyse; Konfiguration in den ersten beiden Dimensionen, mit und ohne überlagerten Multiple Spanning Tree; e) und f) bei ‘Concerto’ in Woche 44 (erklärte Varianz in der ersten Dimension 34,7%, in der zweiten Dimension 15,8%); g) und h) bei ‘Concerto’ in Woche 48 (erklärte Varianz in der ersten Dimension 25,9%, in der zweiten Dimension 14,6%)

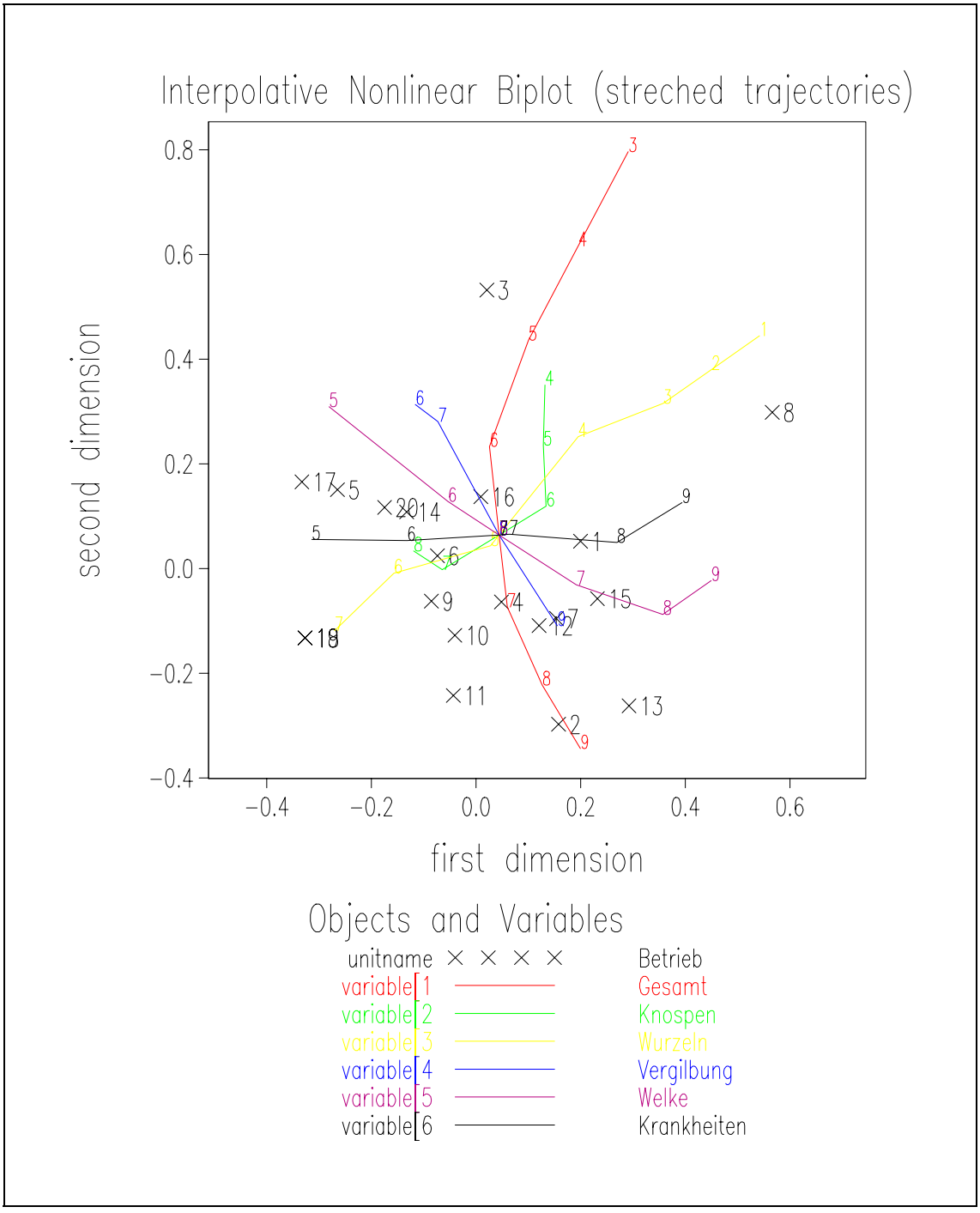


Abbildung A12: Nichtlineare Biplots, ‘Sierra’ Woche 44; Anteil der durch die erste Dimension erklärten Varianz 24,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,5%

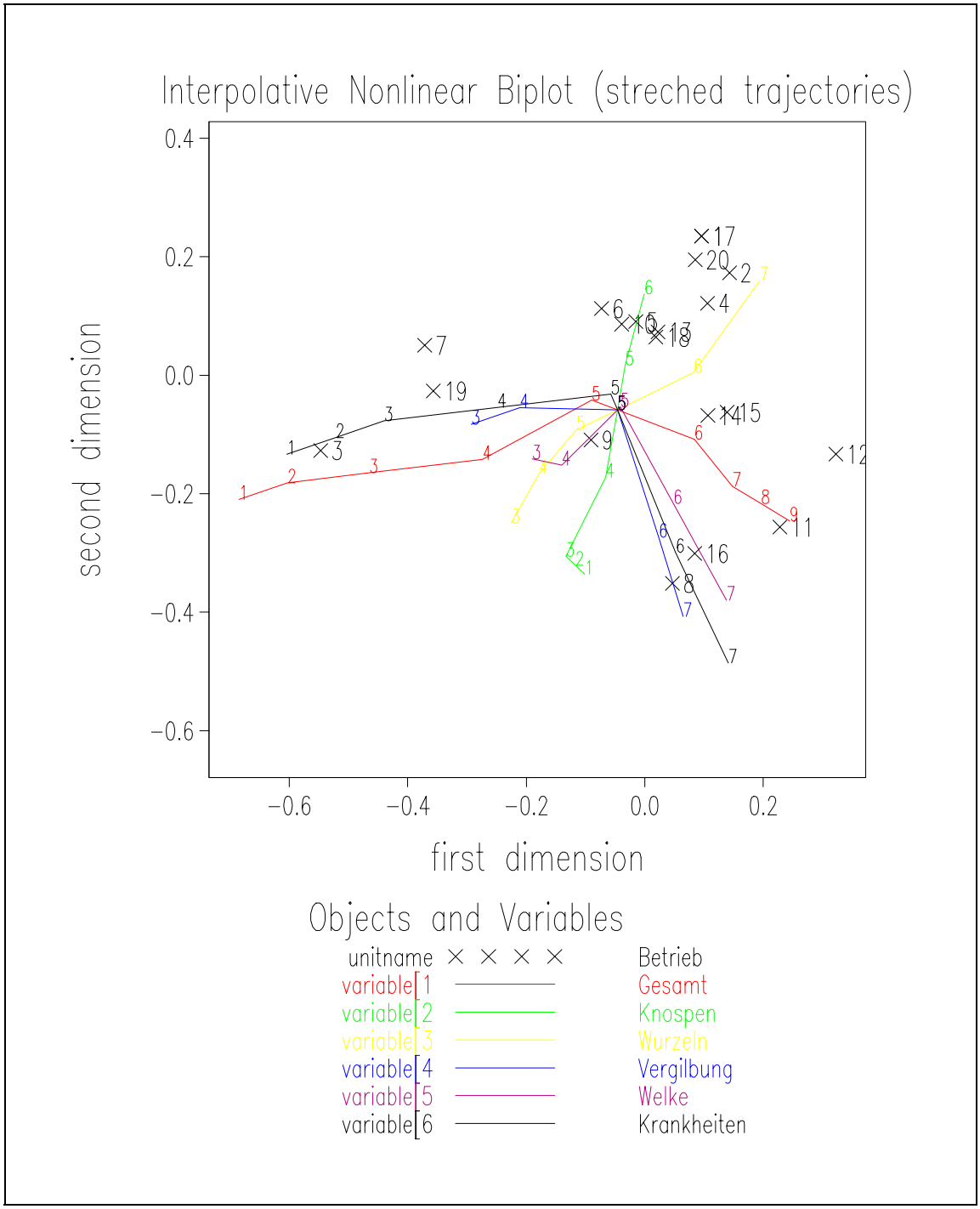


Abbildung A13: Nichtlineare Biplots, ‘Sierra’ Woche 48; Anteil der durch die erste Dimension erklärten Varianz 25,5%, Anteil der durch die zweite Dimension erklärten Varianz 17,4%

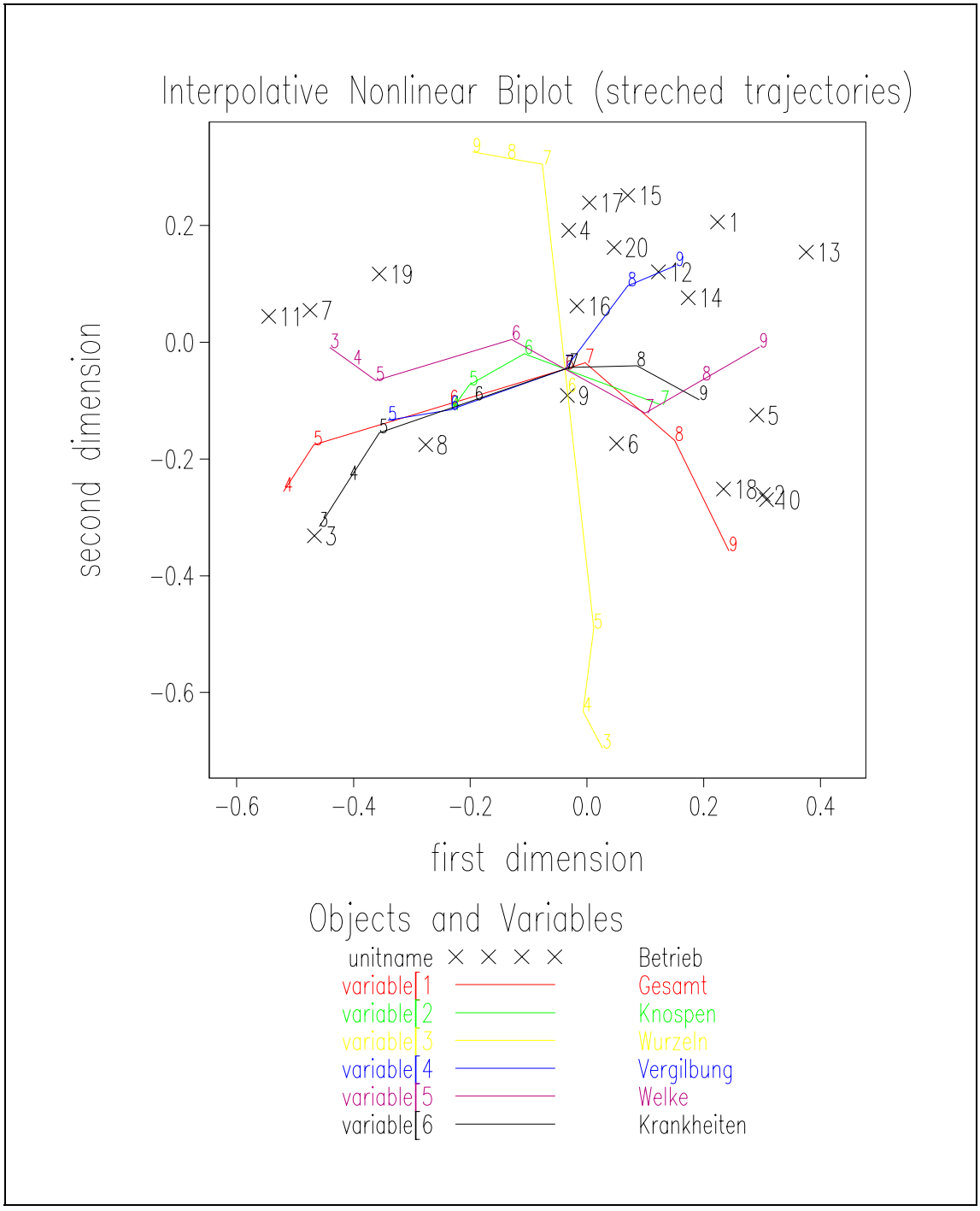


Abbildung A14: Nichtlineare Biplots, ‘Concerto’ Woche 44; Anteil der durch die erste Dimension erklärten Varianz 34,7%, Anteil der durch die zweite Dimension erklärten Varianz 15,8%

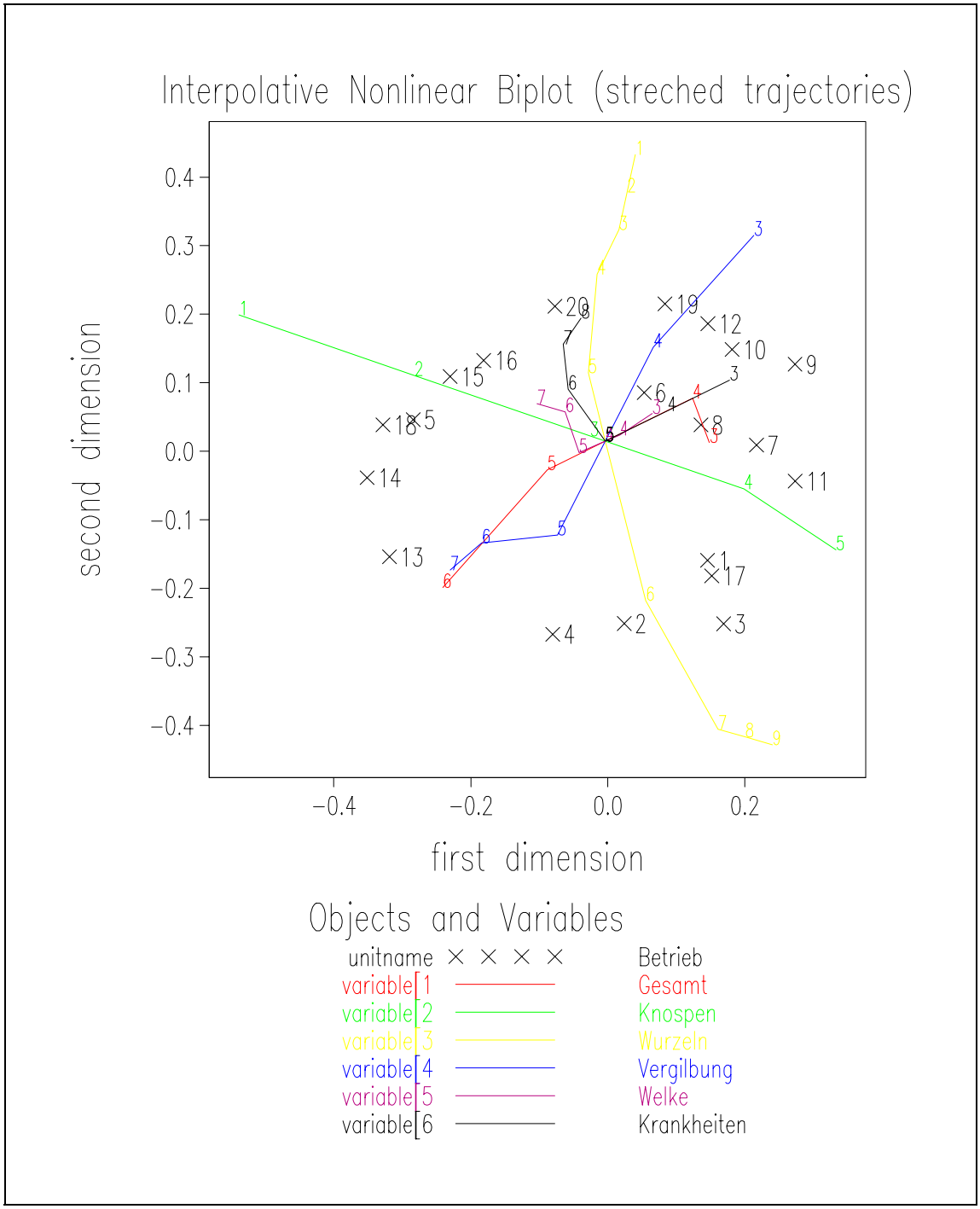


Abbildung A15: Nichtlineare Biplots, ‘Concerto’ Woche 48; Anteil der durch die erste Dimension erklärten Varianz 25,9%, Anteil der durch die zweite Dimension erklärten Varianz 14,6%

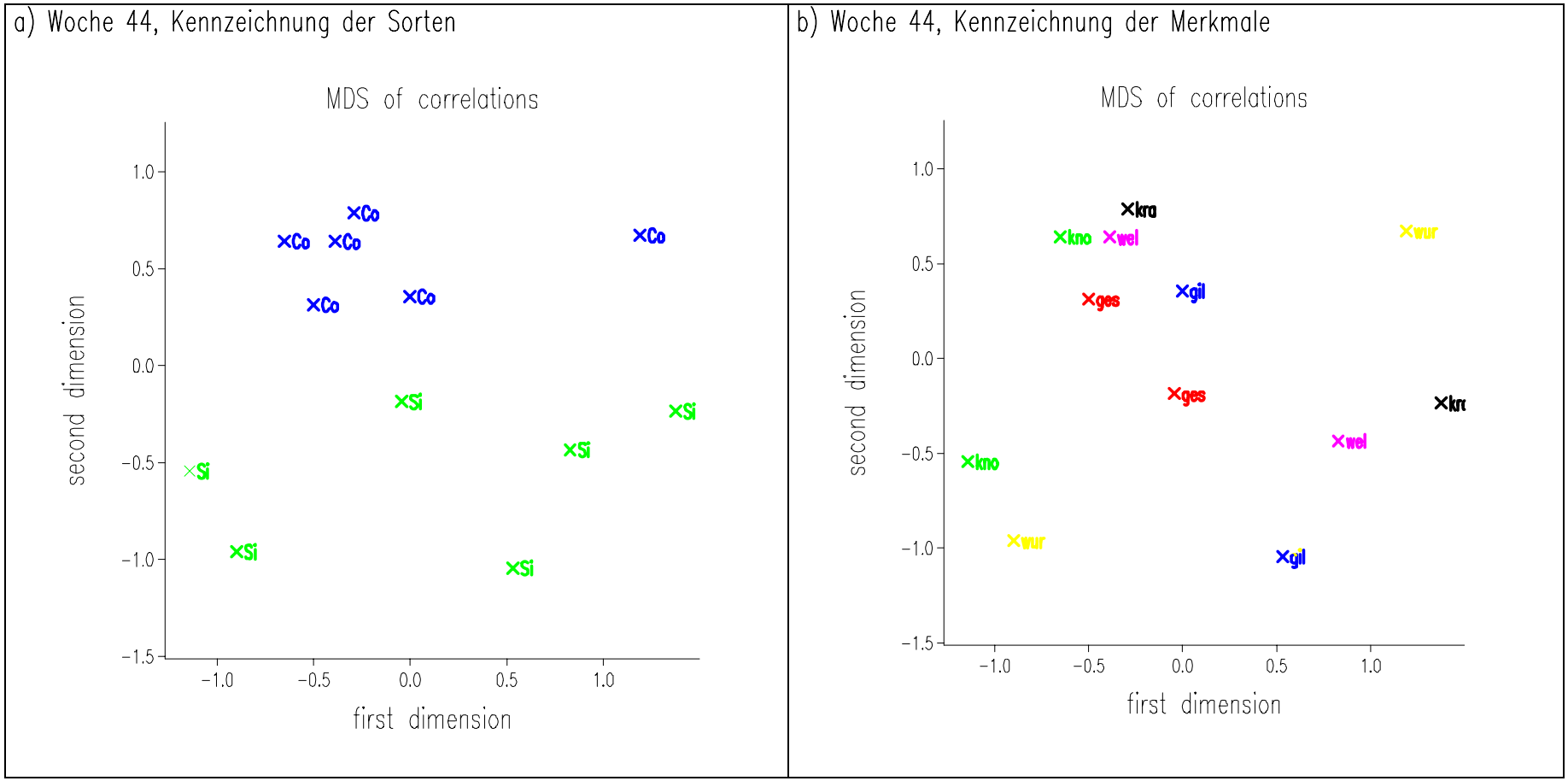


Abbildung A16 a und b: Konfigurationen der ordinalen, mehrdimensionalen Skalierung in den beiden ersten Dimensionen bei Analyse der Spearman Korrelationsmatrix der Boniturwerte für Woche 44, stress-Wert 0,1073

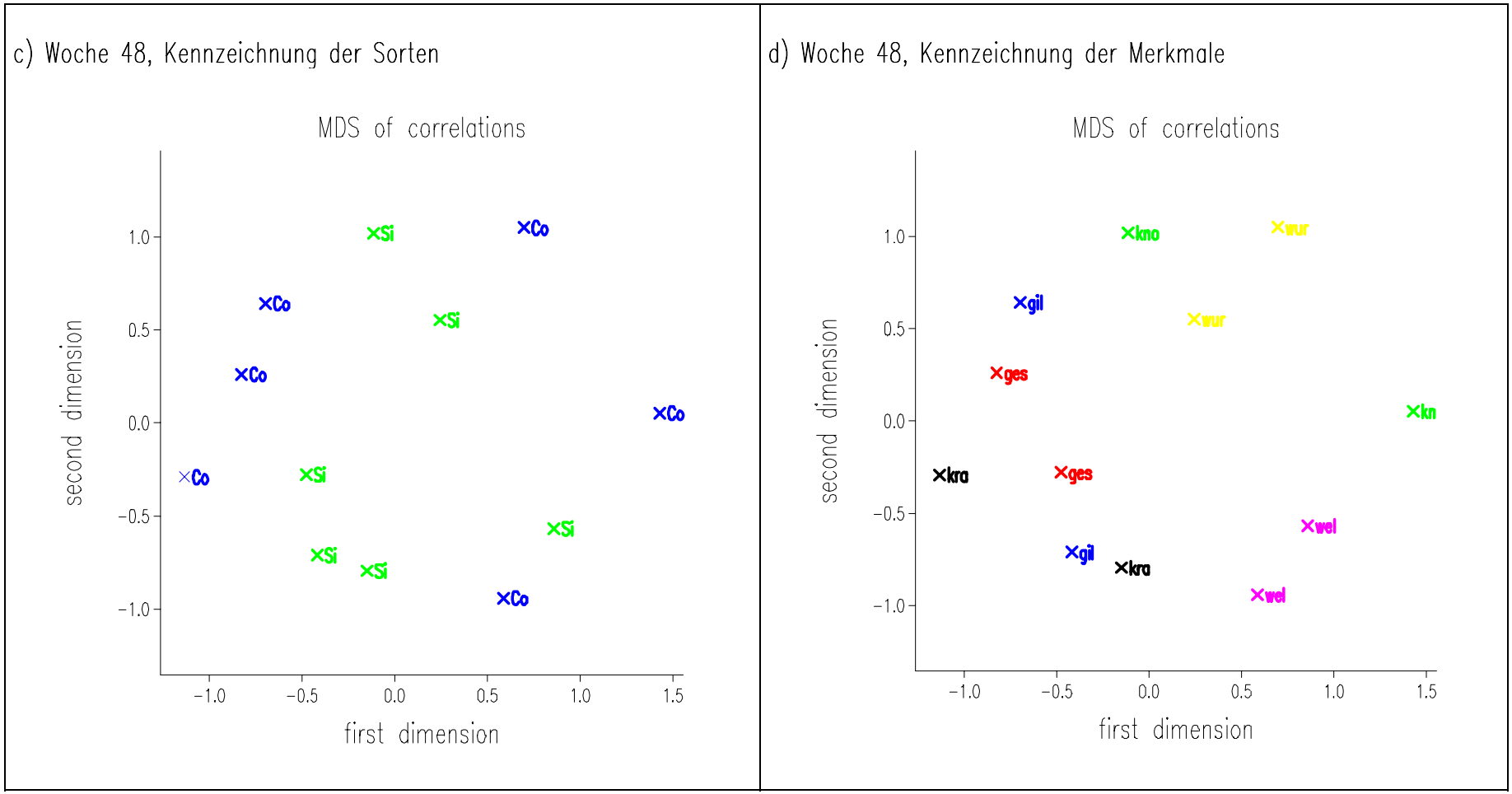


Abbildung A16 c und d: Konfigurationen der ordinalen, mehrdimensionalen Skalierung in den beiden ersten Dimensionen bei Analyse der Spearman Korrelationsmatrix der Boniturwerte für Woche 48, stress-Wert 0,1774

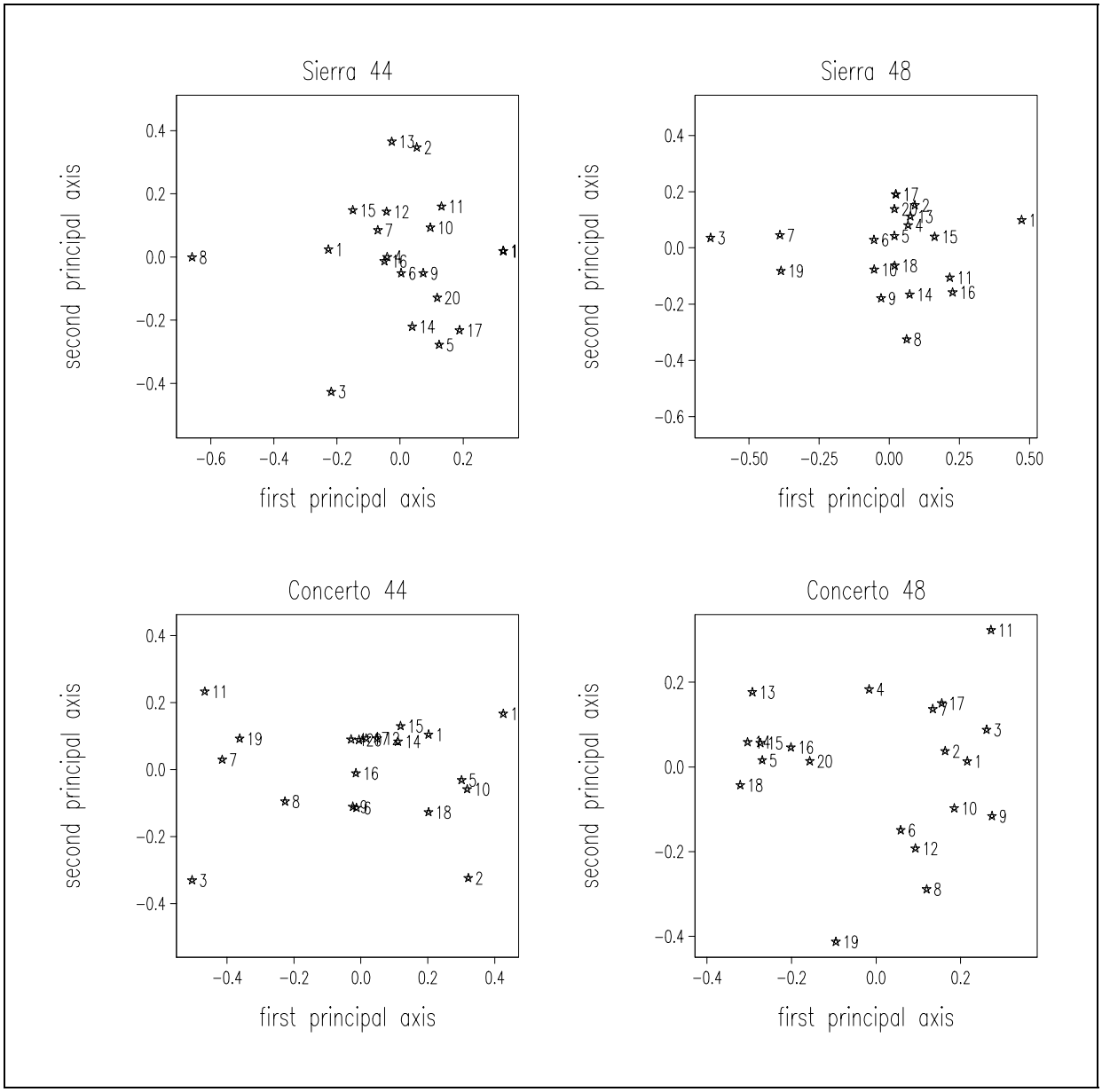


Abbildung A17: Überblick über die Konfigurationen der Korrespondenzanalyse der Qualitätsbonituren in den ersten beiden Dimensionen

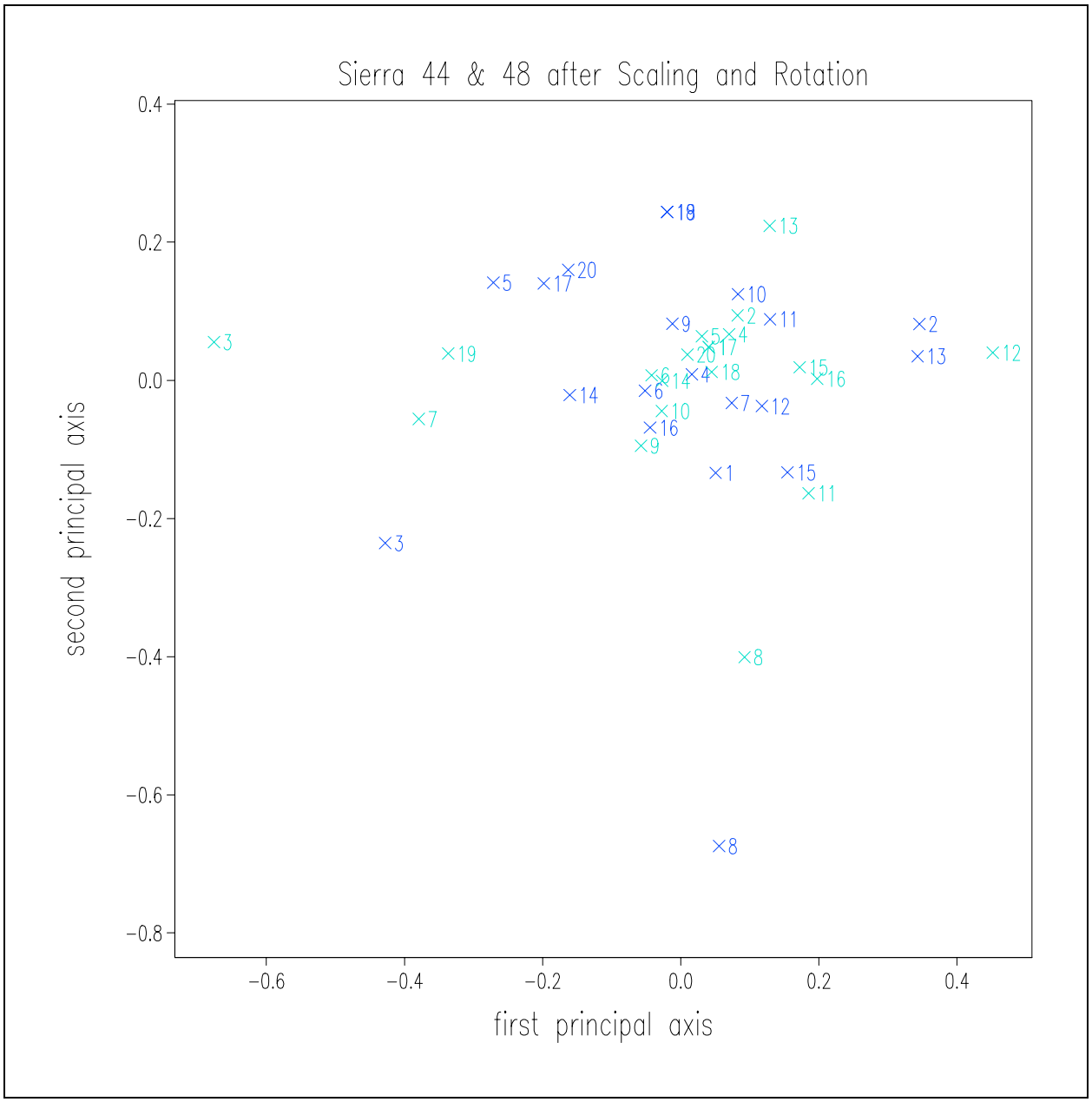


Abbildung A18: Konfigurationen von ‘Sierra’ Woche 44 und ’Sierra’ Woche 48, nach Skalierung und Rotation im Rahmen der Procrustes-Analyse

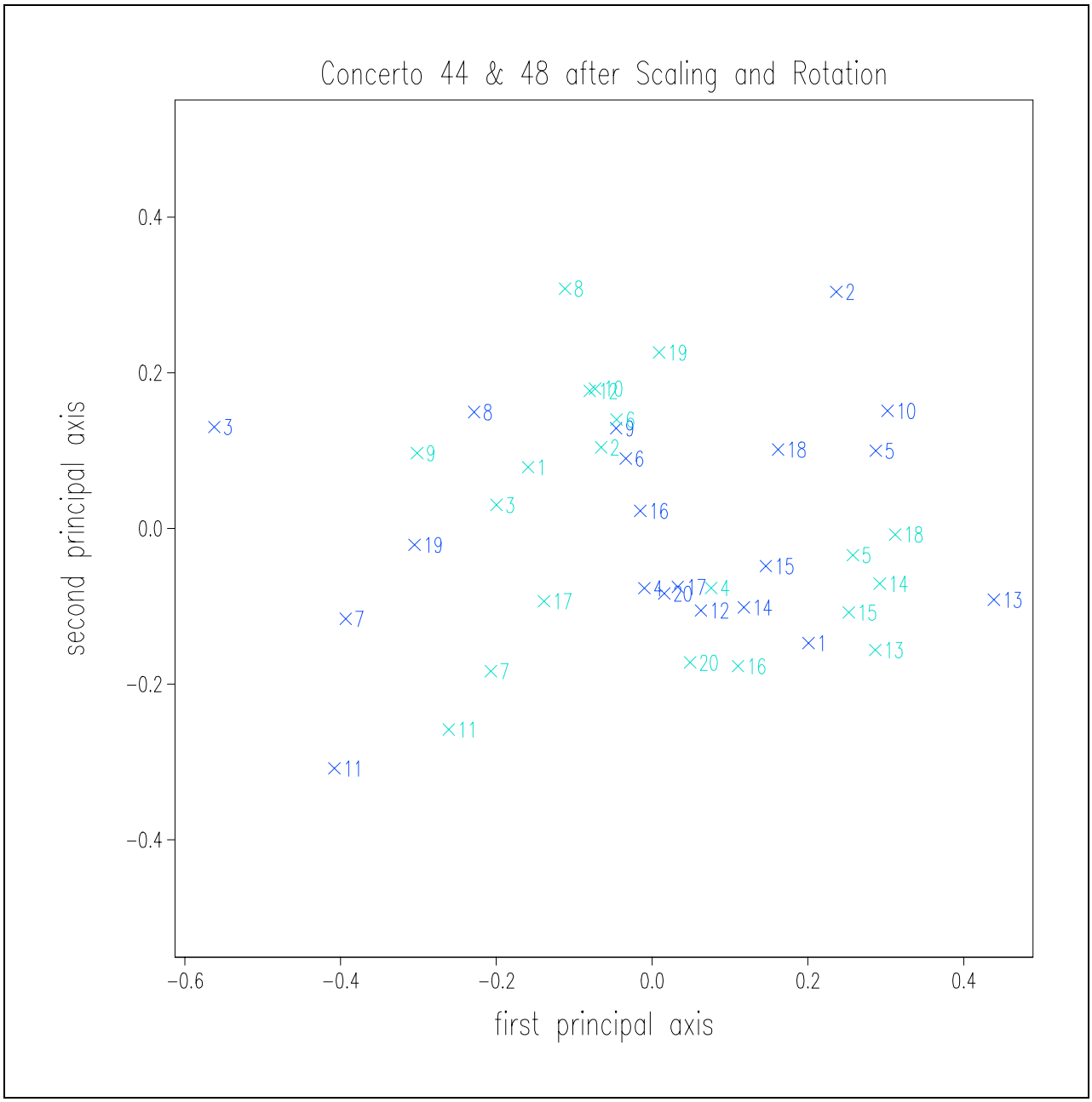


Abbildung A19: Konfigurationen von ‘Concerto’ Woche 44 und ’Concerto’ Woche 48, nach Skalierung und Rotation im Rahmen der Procrustes-Analyse

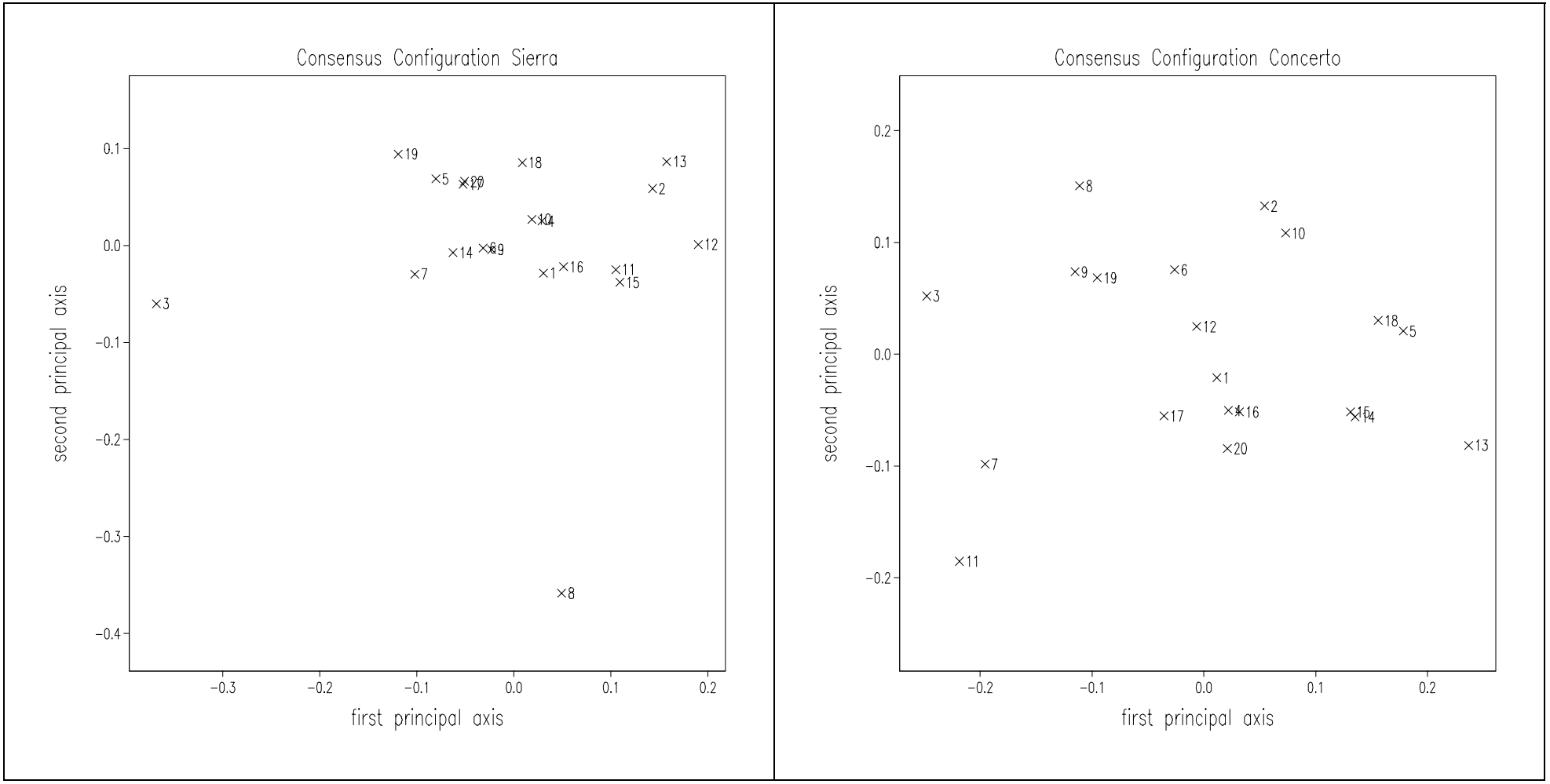


Abbildung A20: Konsens-Konfigurationen der Beurteilungswochen 44 und 48 für ‘Sierra’ und ‘Concerto’; erklärte Varianz durch die erste Dimension bei ‘Sierra’ 39,4%, bei ‘Concerto’ 43,9%, durch die zweite Dimension bei ‘Sierra’ 24,3%, bei ‘Concerto’ 18,6%

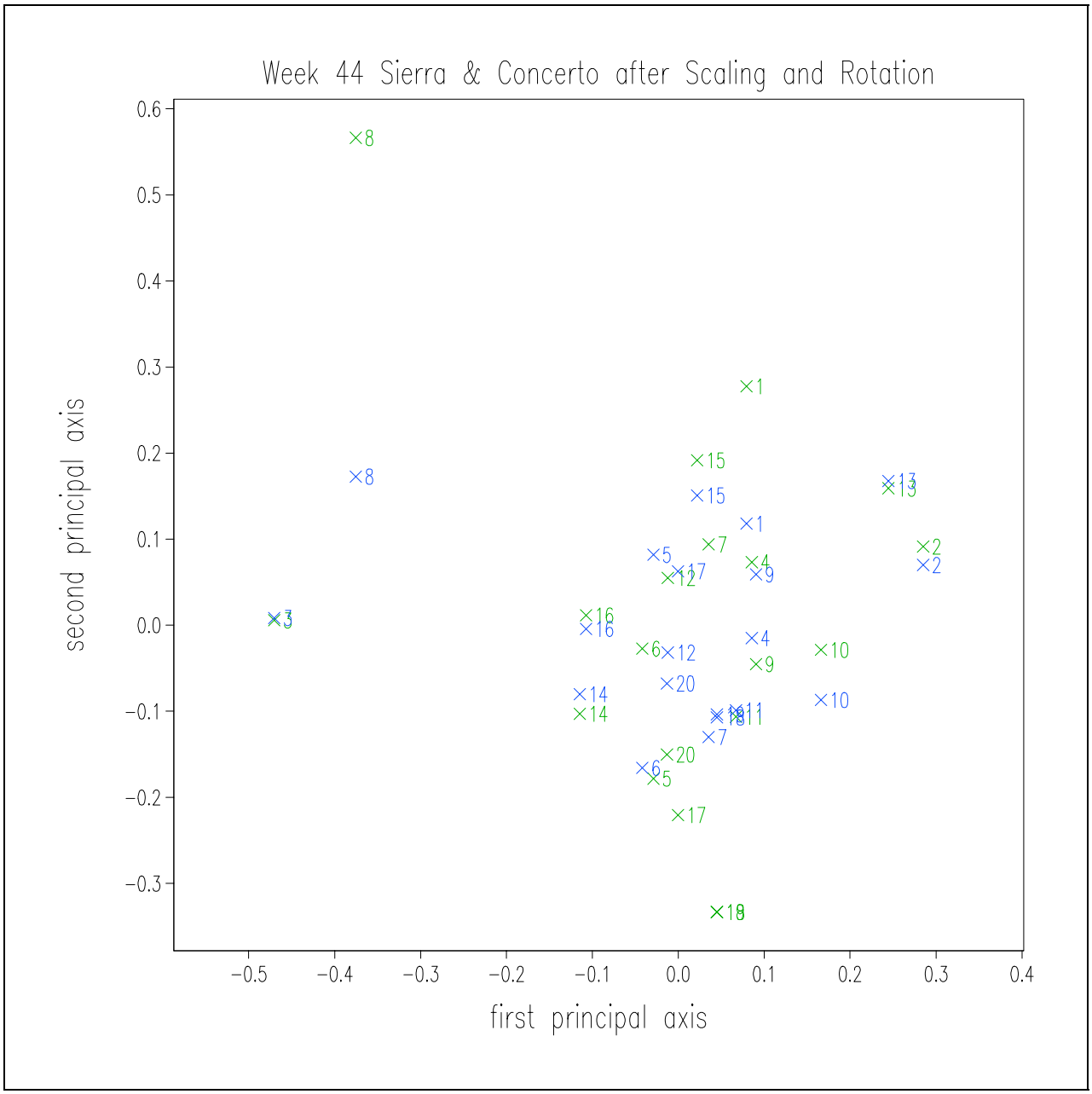


Abbildung A21: Konfigurationen von 'Sierra' Woche 44 und 'Concerto' Woche 44, nach Skalierung und Rotation im Rahmen der Prokrustes-Analyse

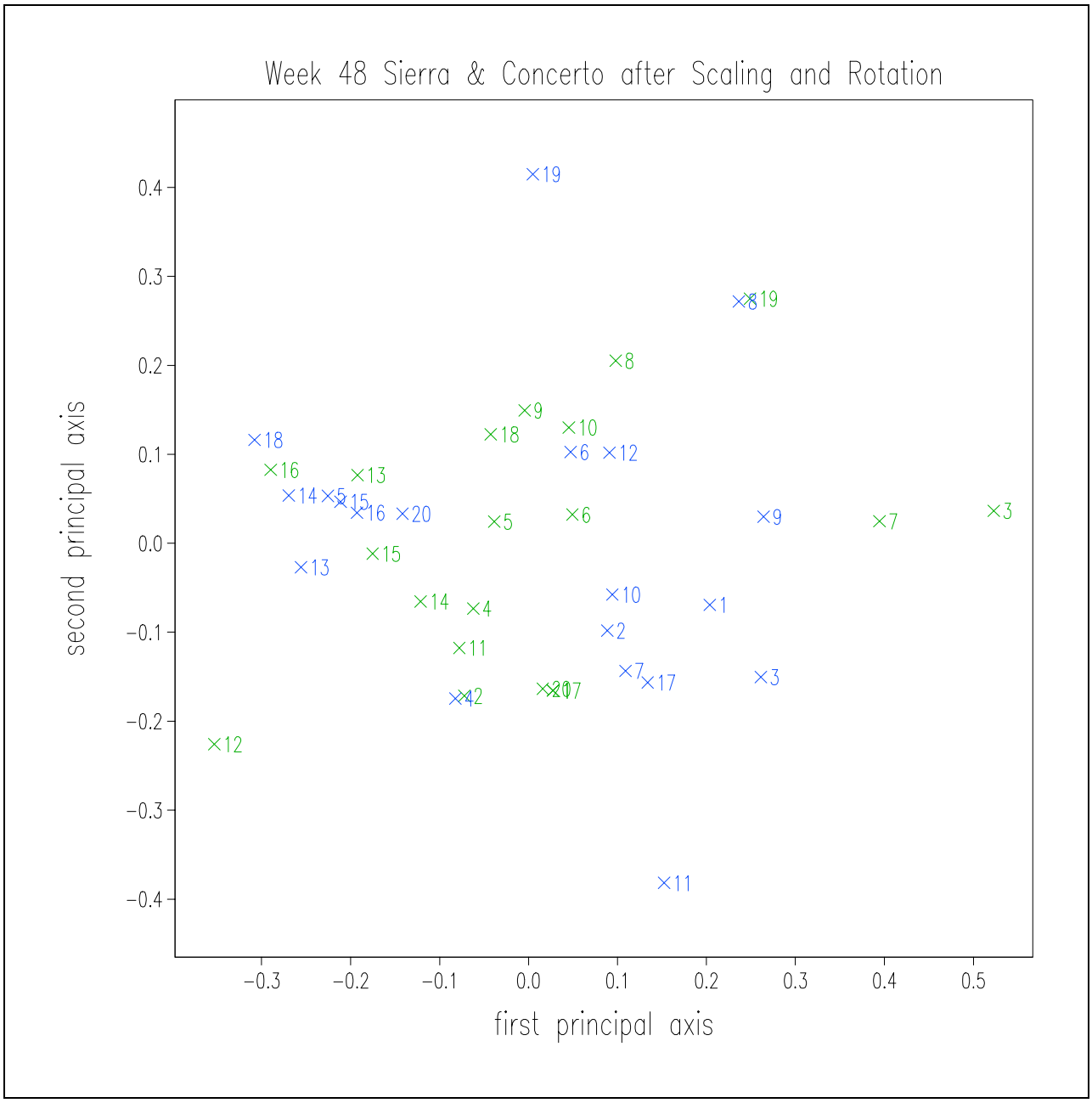


Abbildung A22: Konfigurationen von 'Sierra' Woche 48 und 'Concerto' Woche 48, nach Skalierung und Rotation im Rahmen der Prokrustes-Analyse

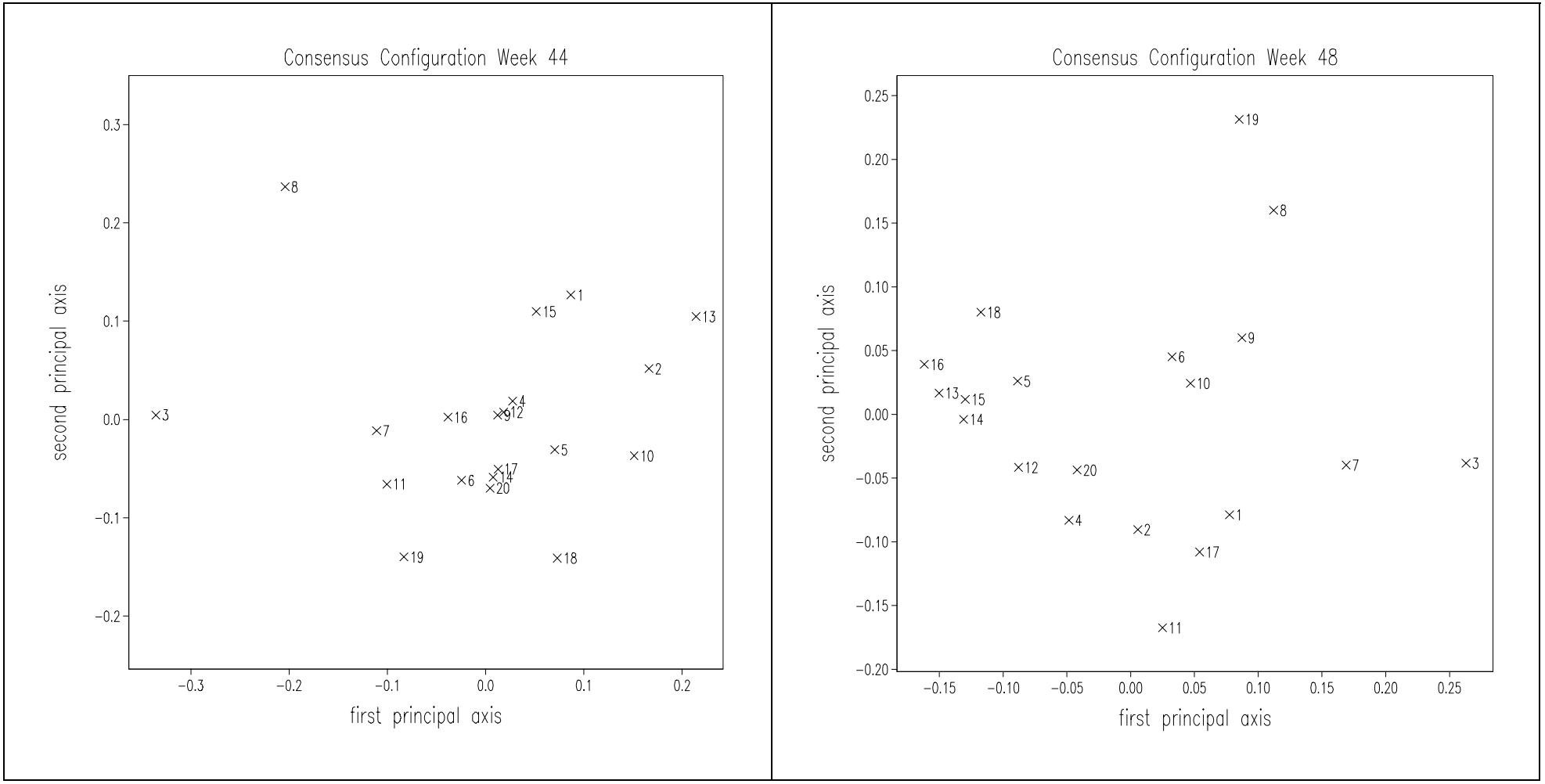


Abbildung A23: Konsens-Konfigurationen der Beurteilungswochen 44 und 48; erklärte Varianz durch die erste Dimension in Woche 44 41,3%, in Woche 48 36,3%, durch die zweite Dimension in Woche 44 21,7%, in Woche 48 23,1%

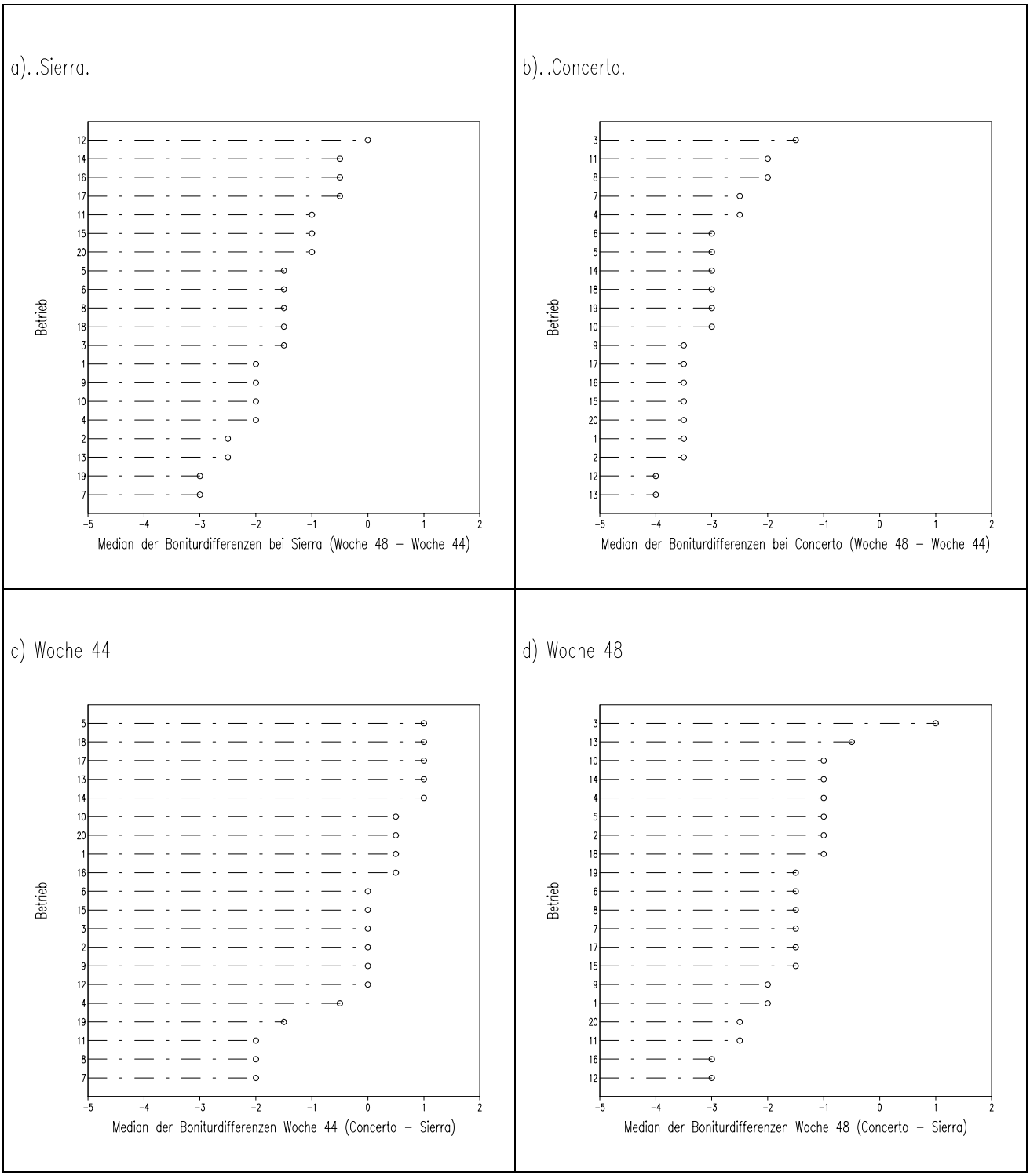


Abbildung A24 a,b,c,d: Dotplots der Boniturdifferenzen zwischen ‘Sierra’ in Woche 48 und 44 (a)); ‘Concerto’ in Woche 48 und 44 (b)); in Woche 44 zwischen ‘Concerto’ und ‘Sierra’ (c)); in Woche 48 zwischen ‘Concerto’ und ‘Sierra’ (d))

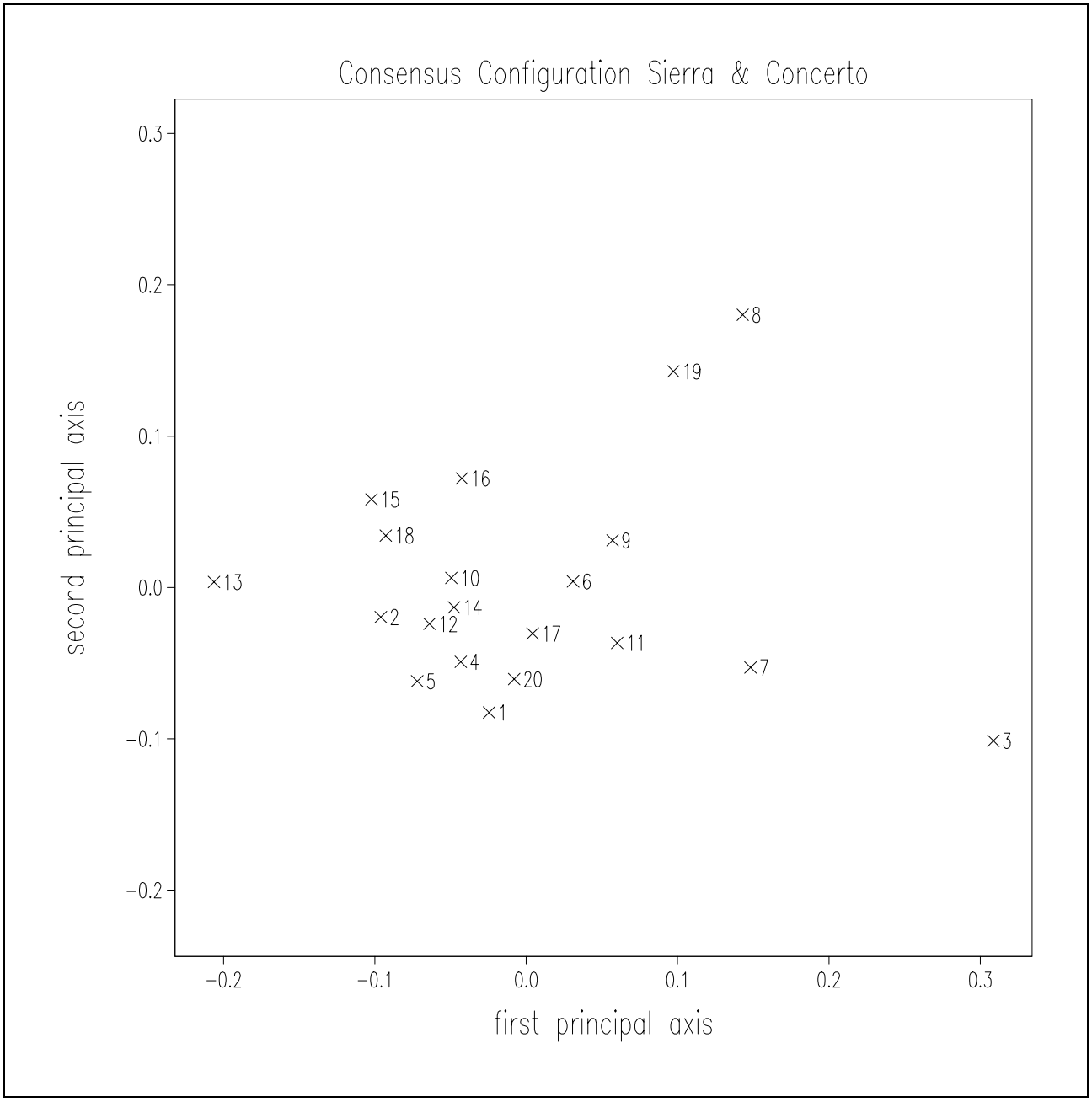


Abbildung A25: Konsens-Konfiguration nach Prokrustes Analyse für ‘Sierra’ und ‘Concerto’, Woche 44 und 48; erklärte Varianz in der ersten Dimension 41,7%, in der zweiten Dimension 16,6%

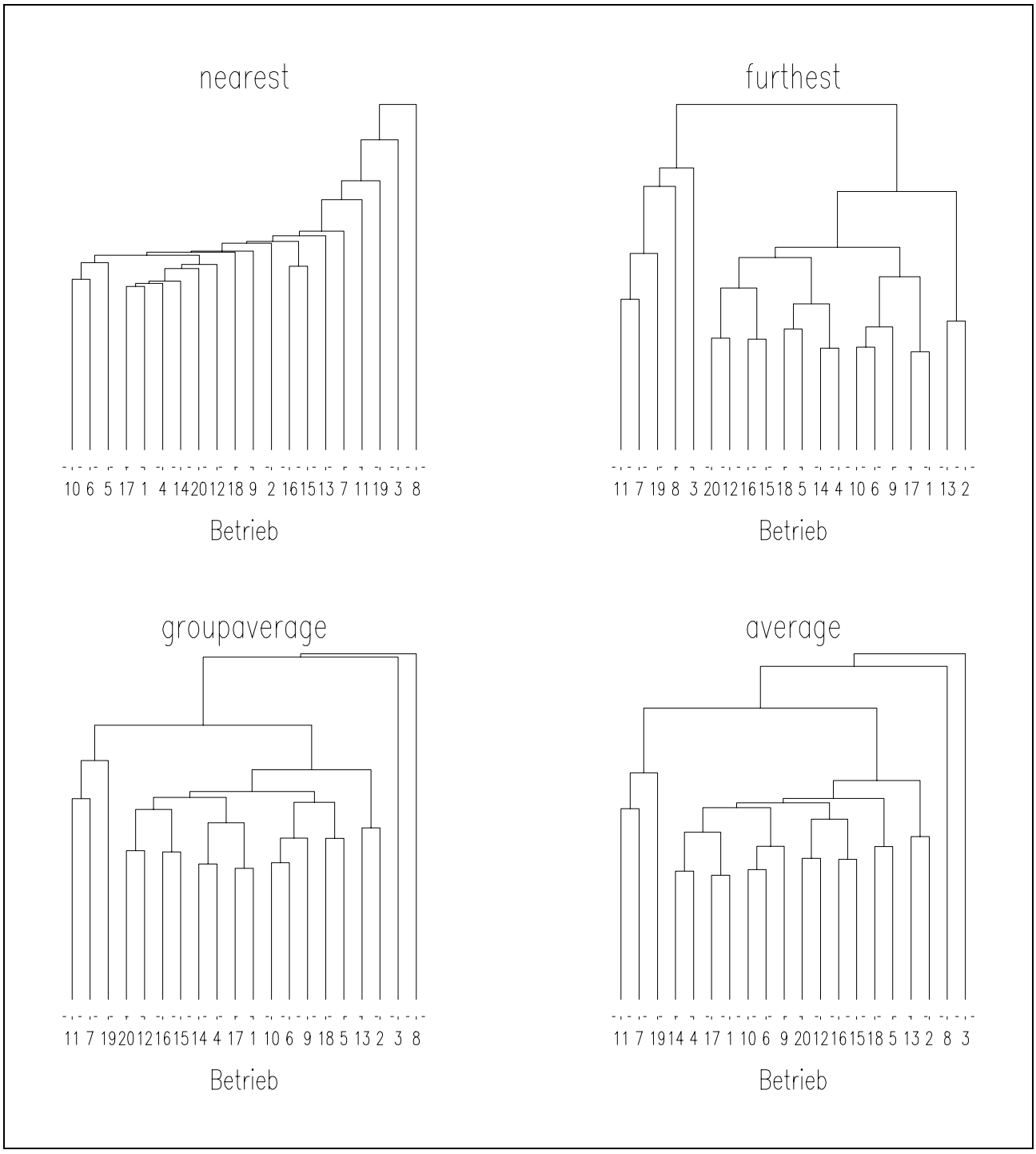


Abbildung A26: Dendrogramme unterschiedlicher Clusteralgorithmen bei Analyse aller Boniturwerte der Woche 44 und 48 bei den Sorten ‘Sierra’ und ‘Concerto’

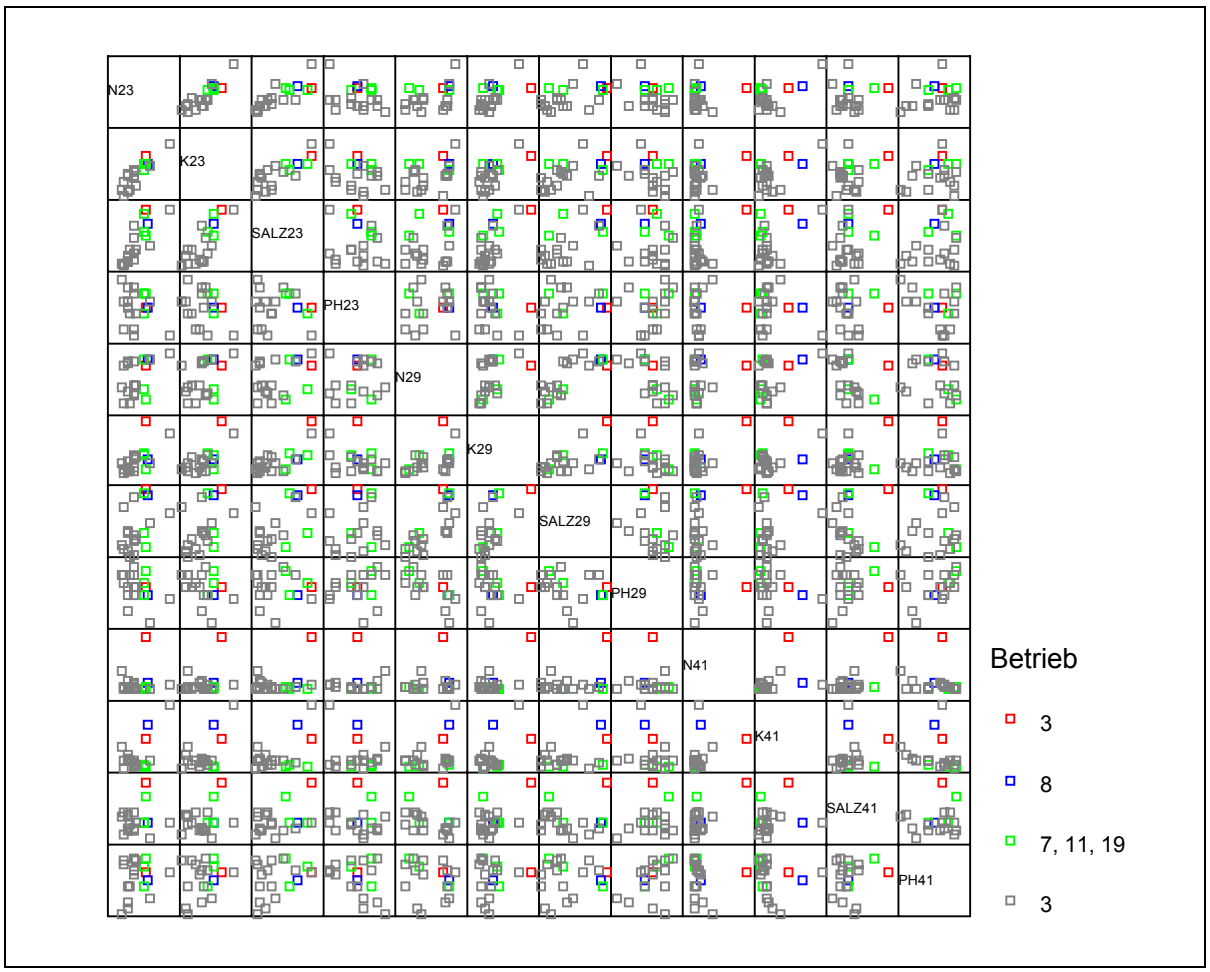


Abbildung A27: Scatterplotmatrix der Substratanalysewerte

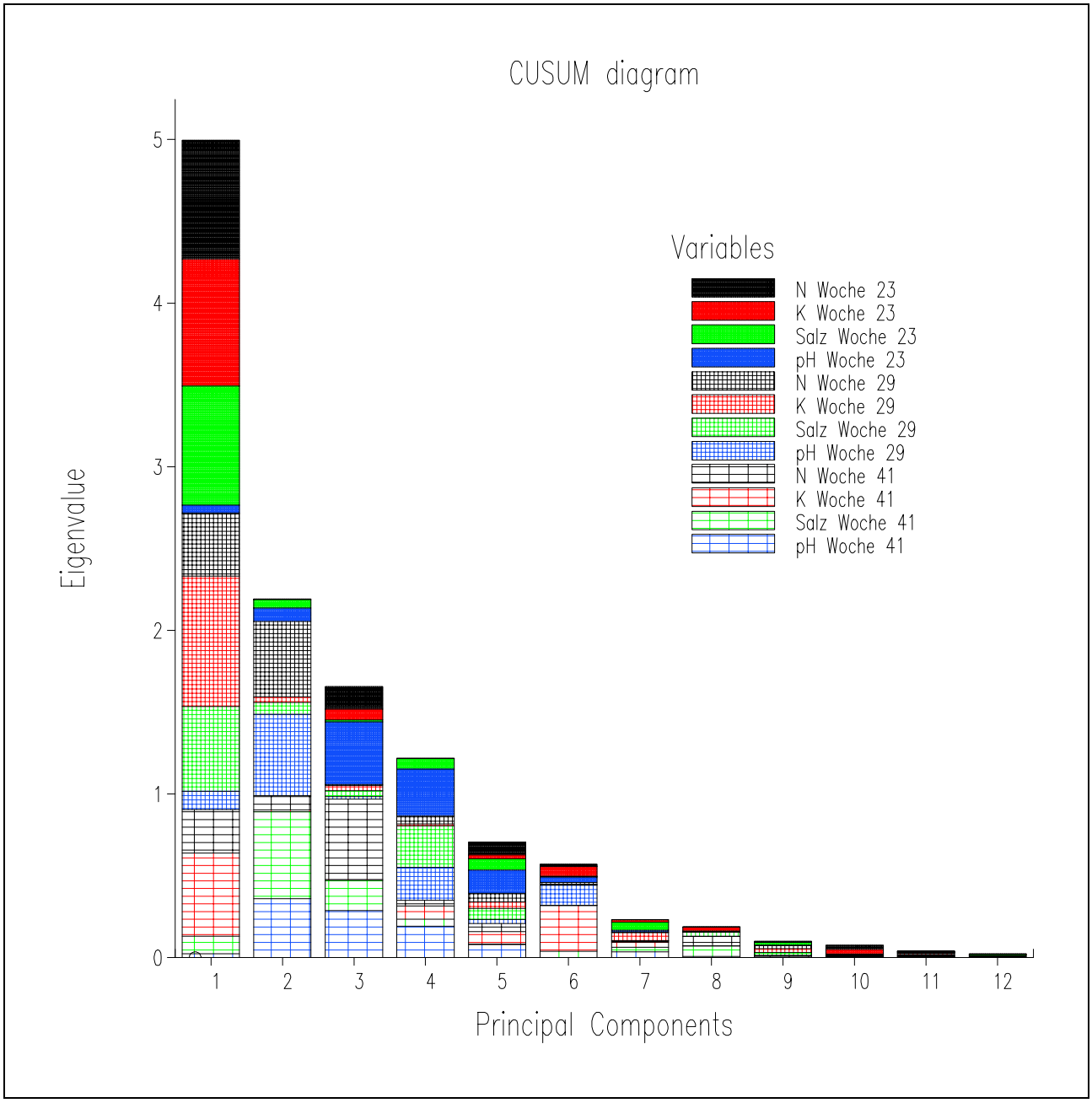


Abbildung A28: CUSUM Diagramm nach Hauptkomponentenanalyse der Substratanalysewerte

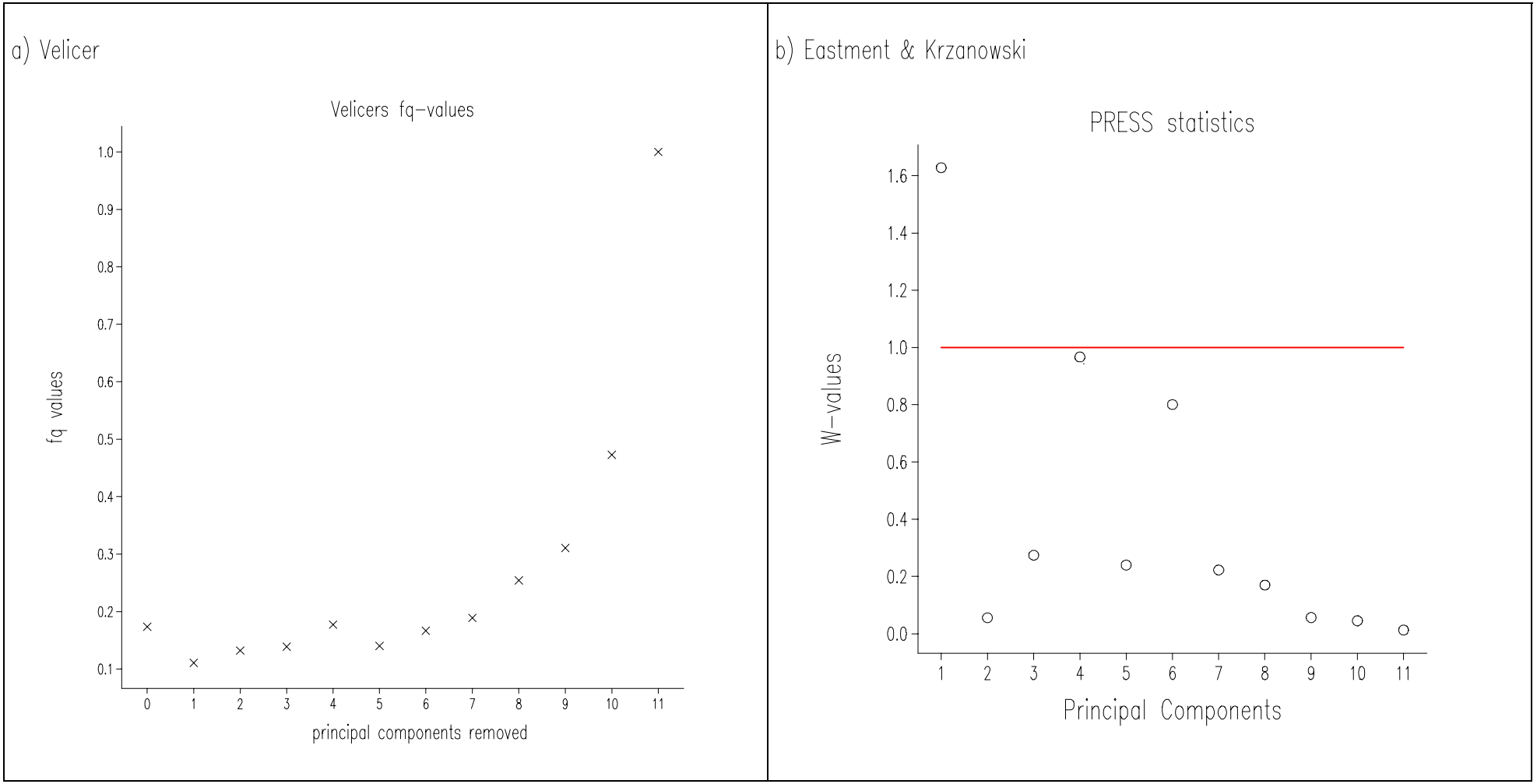


Abbildung A29 a und b: Bestimmung der Anzahl ‘wesentlicher’ Hauptkomponenten nach VELICER, 1976 (a)) und EASTMENT & KRZANOWSKI, 1982 (b)) nach Hauptkomponentenanalyse der Substratanalysewerte

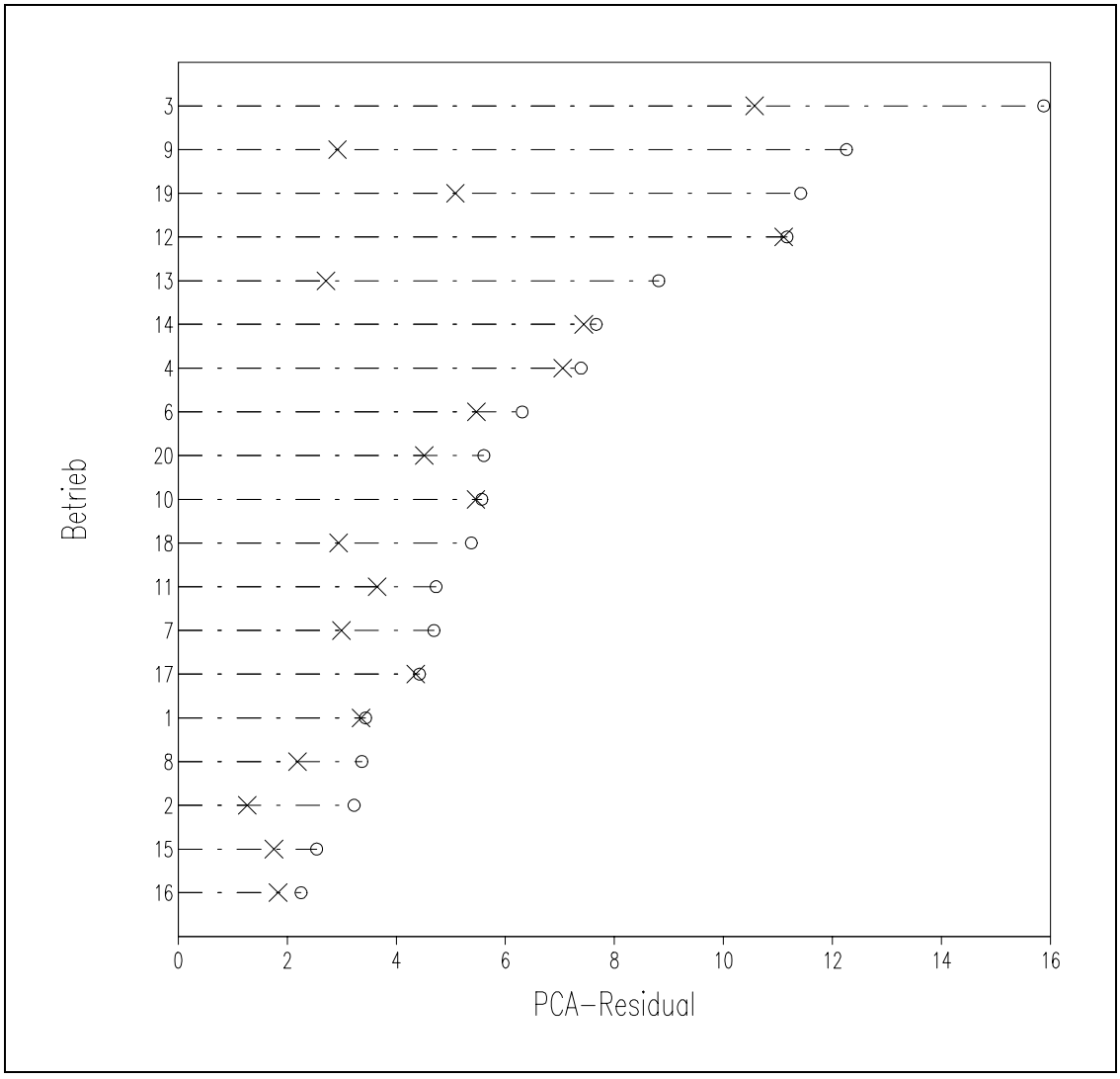


Abbildung A30: Dotplot der Hauptkomponenten-Residuen nach Hauptkomponentenanalyse der Substratanalysewerte und Betrachtung von einer Dimension (Kreis) beziehungsweise von zwei Dimensionen (Kreuz)

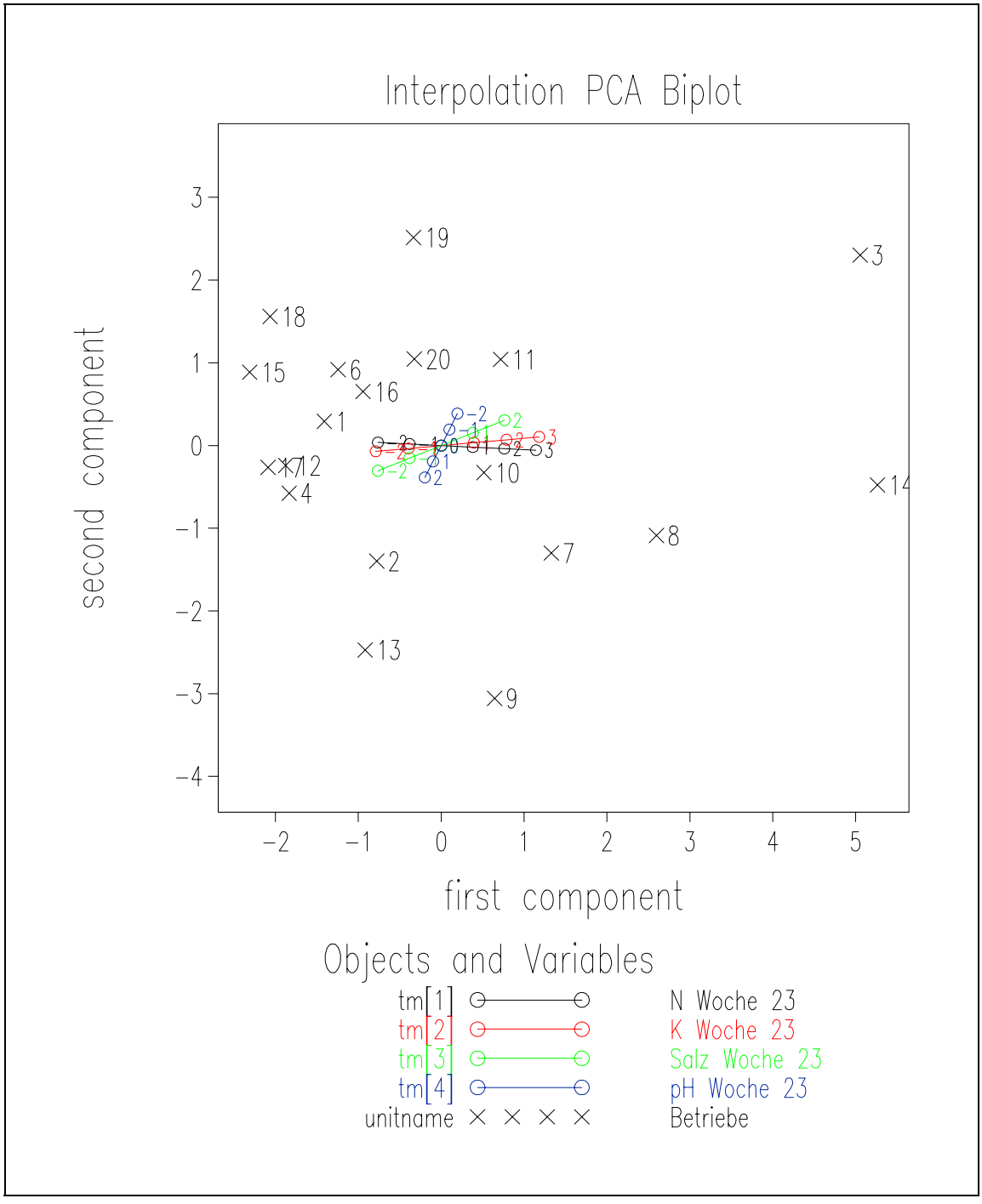


Abbildung A31: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 23 mit Interpolationsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

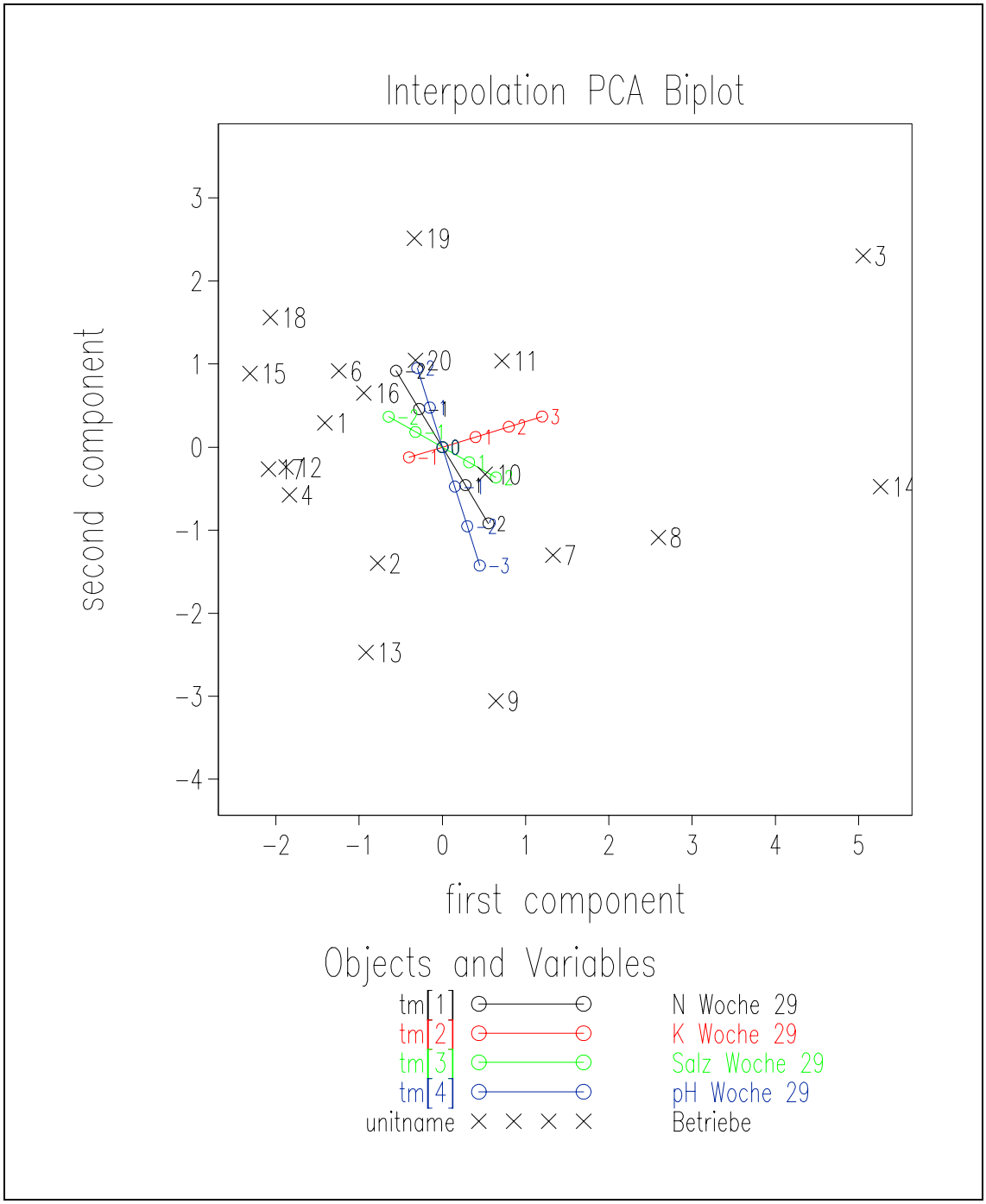


Abbildung A32: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 29 mit Interpolationsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

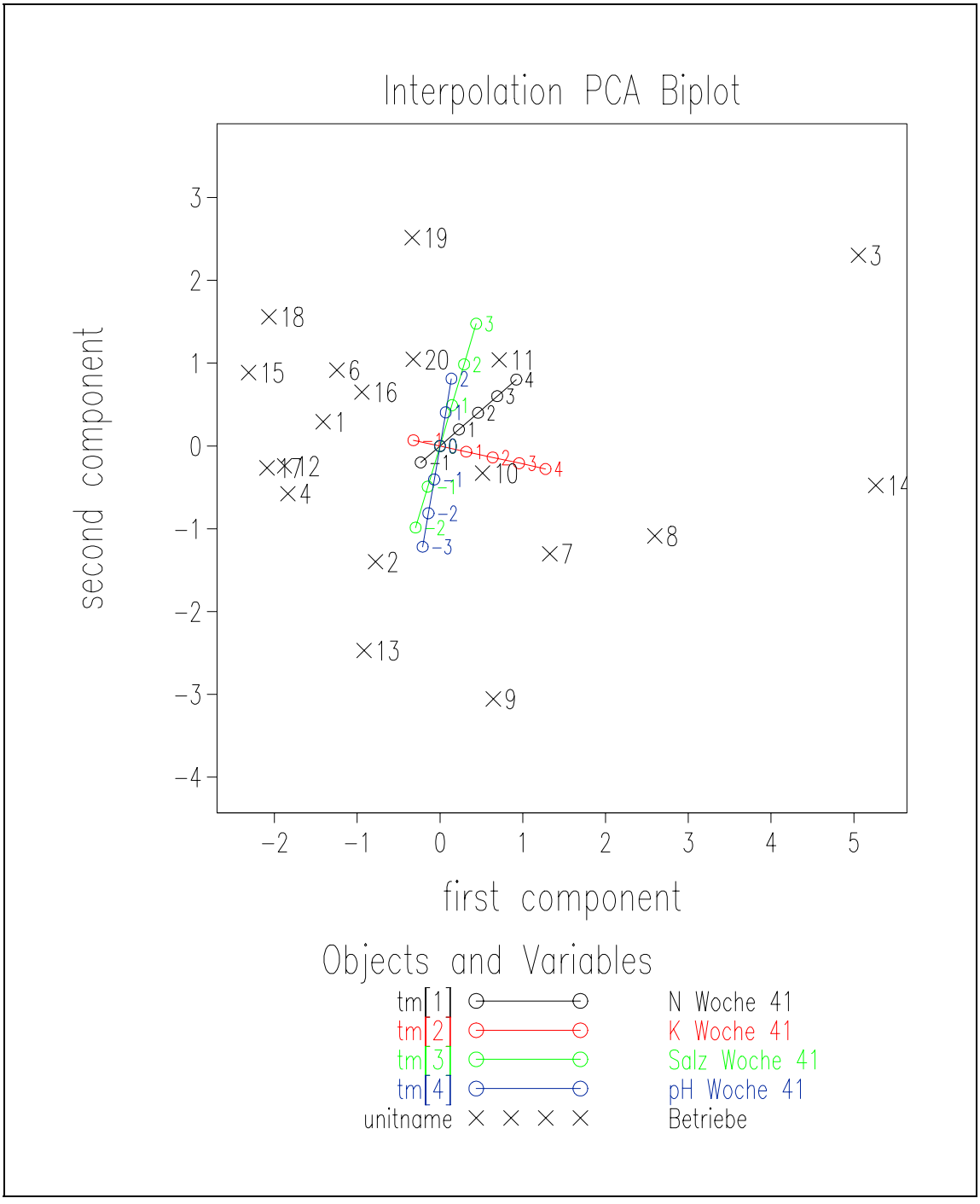


Abbildung A33: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 41 mit Interpolationsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

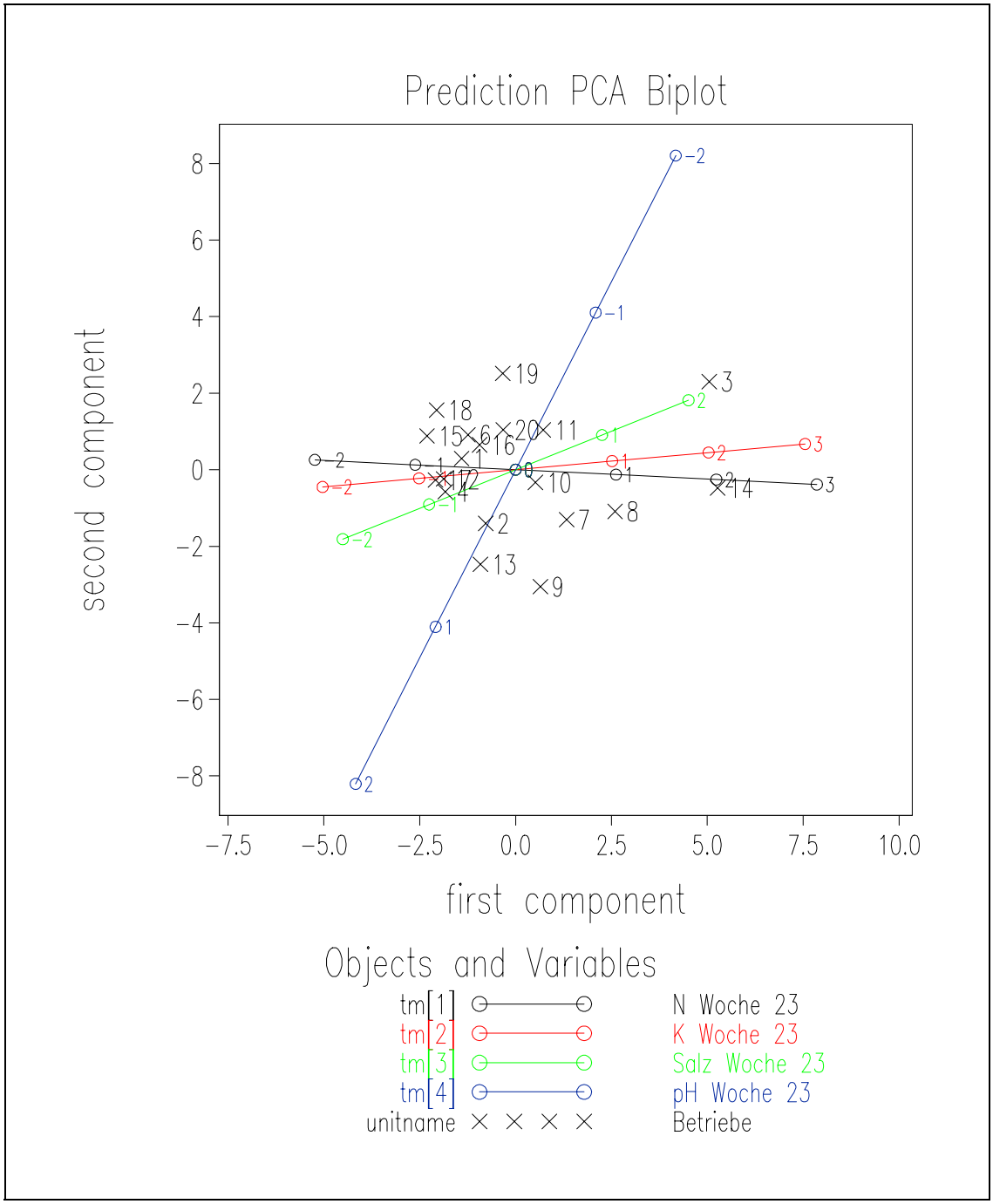


Abbildung A34: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 23 mit Prediktionsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

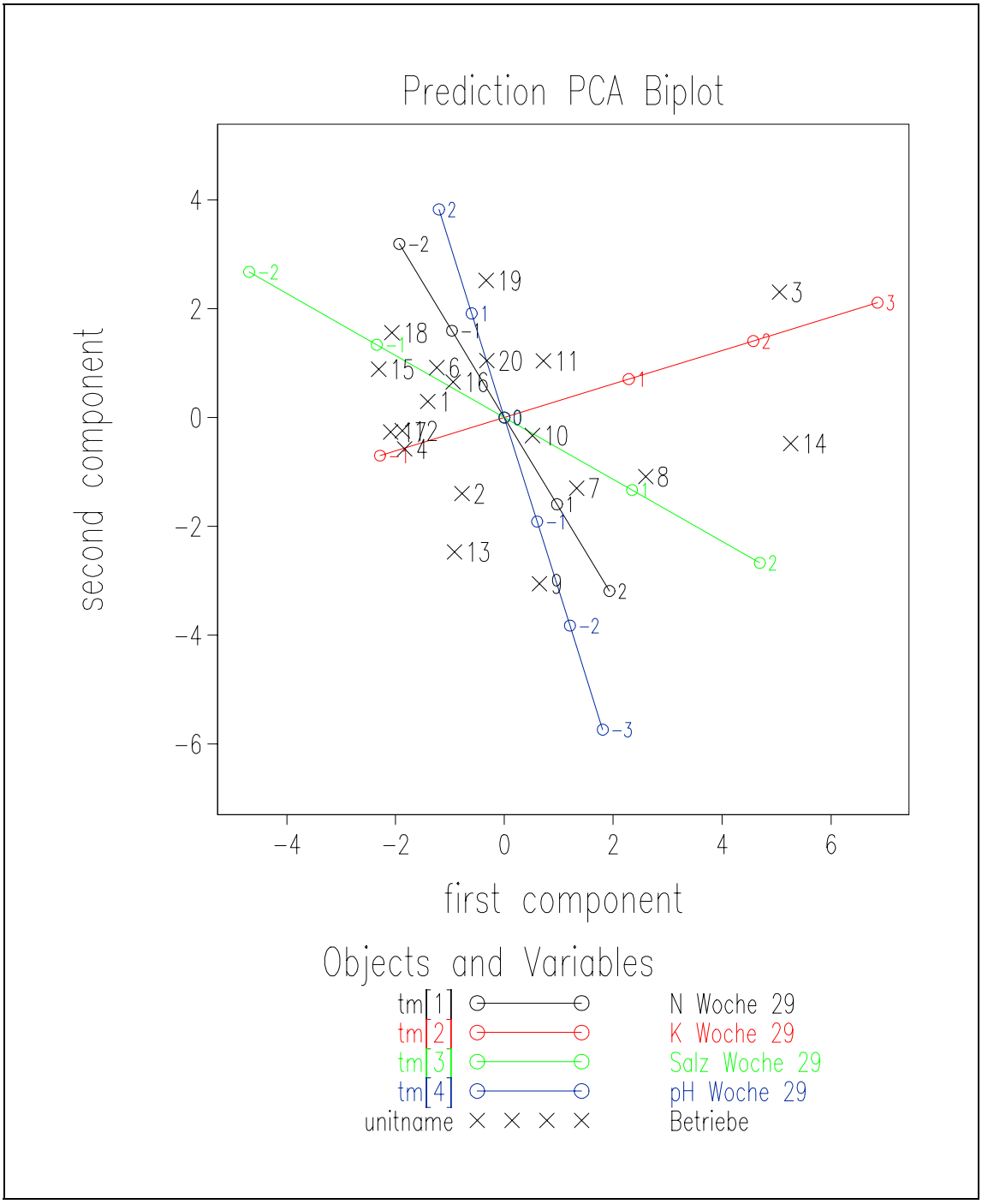


Abbildung A35: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 29 mit Prediktionsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

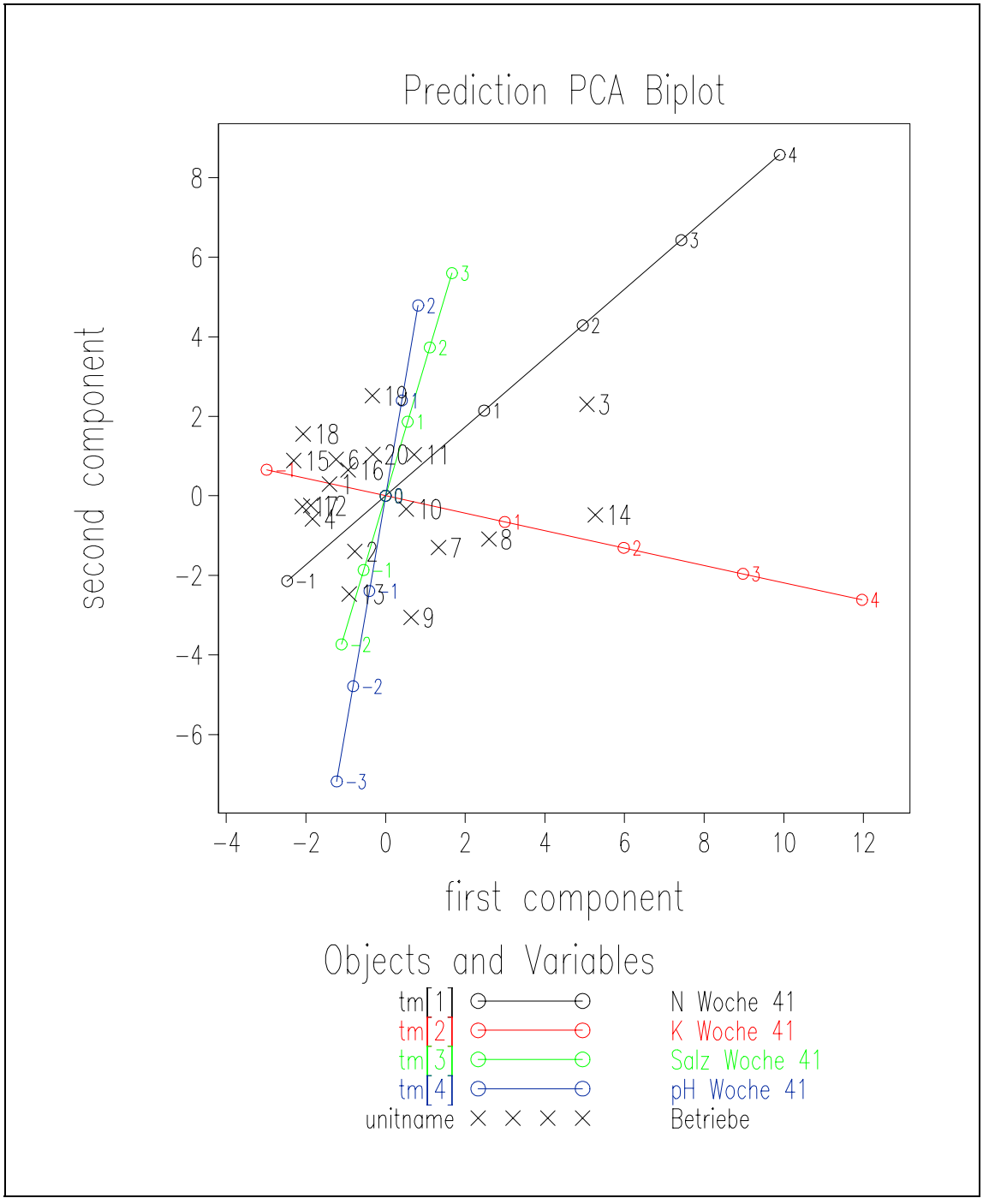


Abbildung A36: Hauptkomponenten-Biplots der Substratanalysewerte in Woche 41 mit Prediktionsmarkern; Anteil der durch die erste Dimension erklärten Varianz 41,6%, Anteil der durch die zweite Dimension erklärten Varianz 18,3%

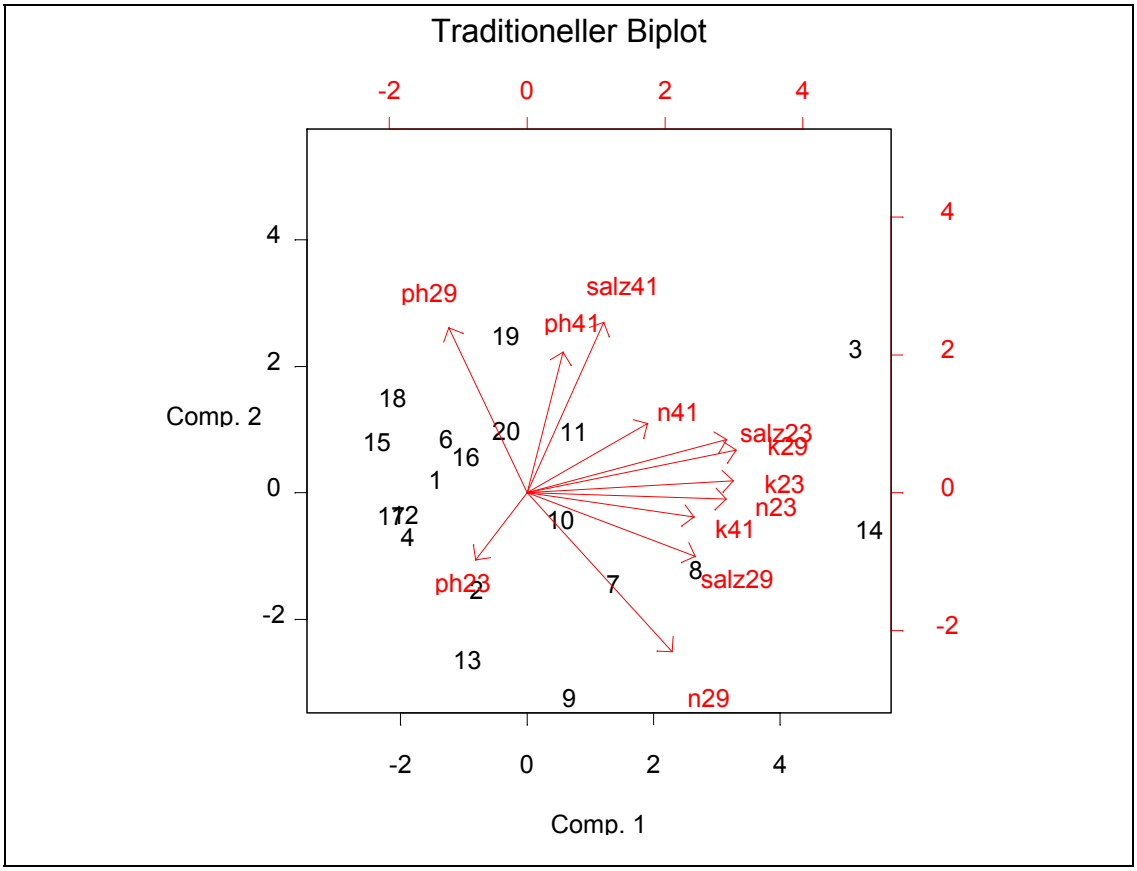


Abbildung A37: Herkömmliche Biplot-Darstellung der Substratanalysewerte

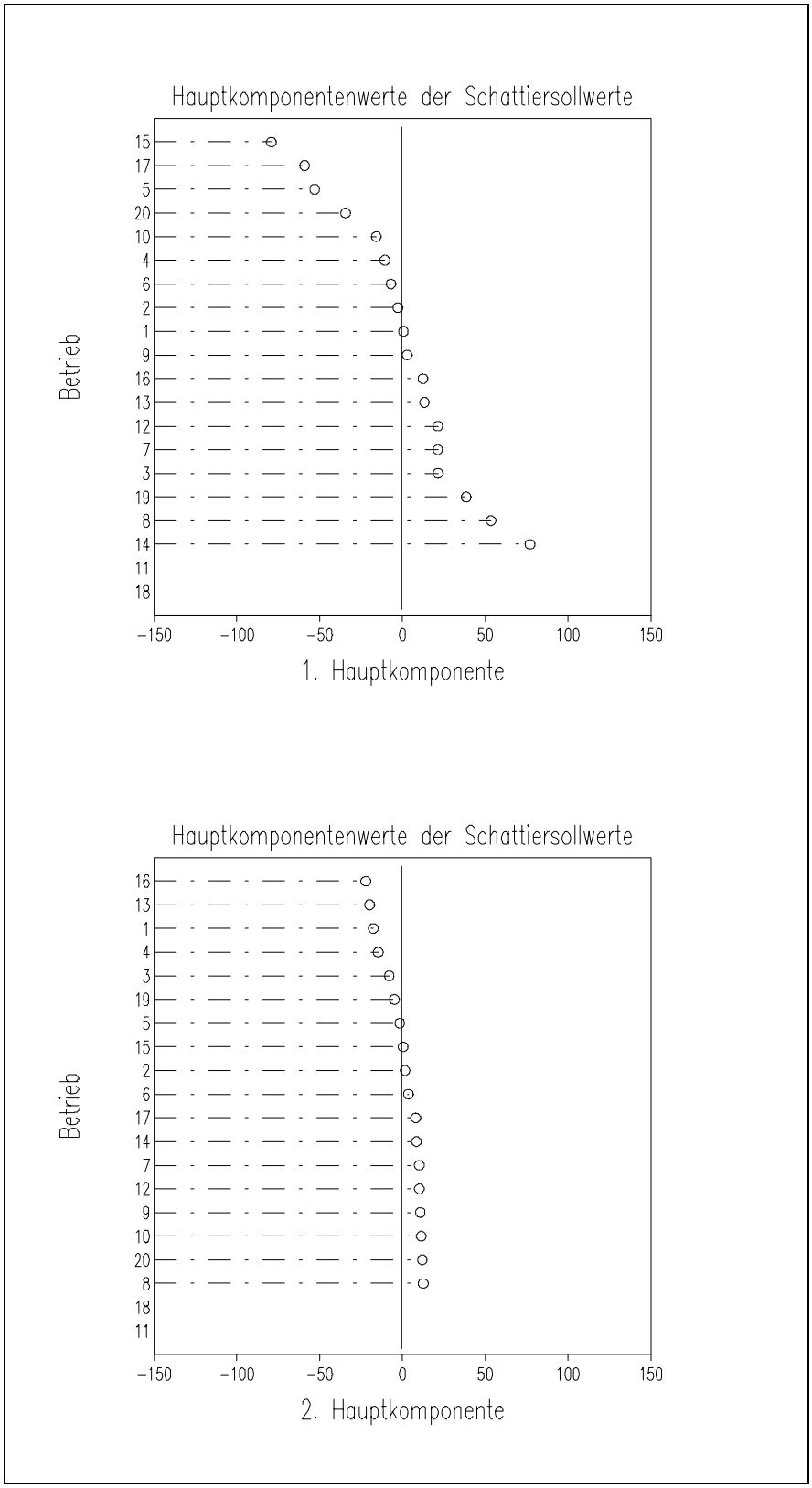


Abbildung A38 : Dotplots der Hauptkomponentenwerte nach Hauptkomponentenanalyse der Schattiersollwerte (ohne Betrieb 11 und 18, da keine Angaben), a) der ersten Hauptkomponente, b) der zweiten Hauptkomponente

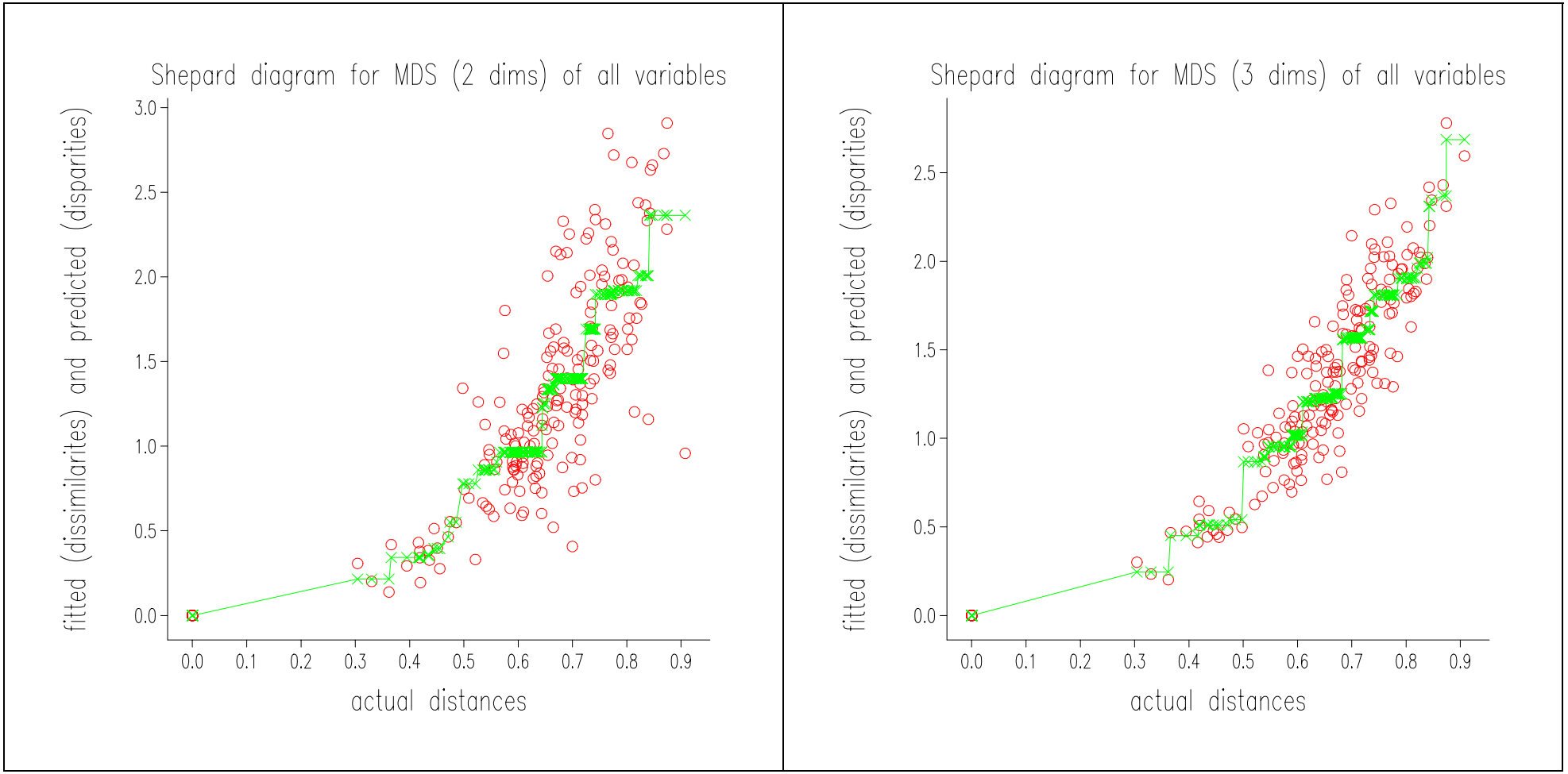


Abbildung A39: Shepard-Plots nach ordinaler mehrdimensionaler Skalierung bei Skalierung in zwei (2 dims) und drei (3 dims) Dimensionen

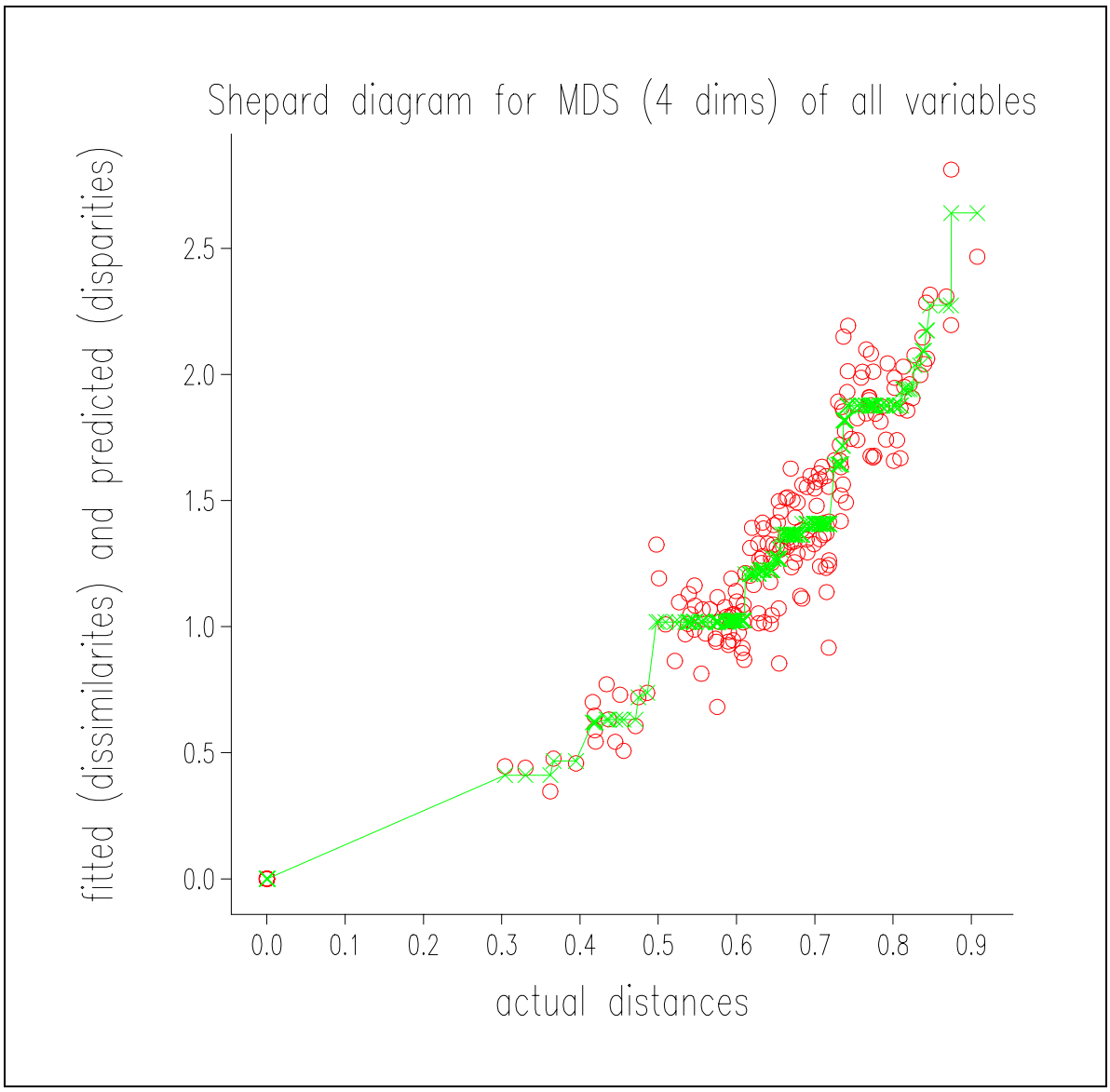


Abbildung A40: Shepard-Plot nach ordinaler mehrdimensionaler Skalierung in vier (4 dims) Dimensionen

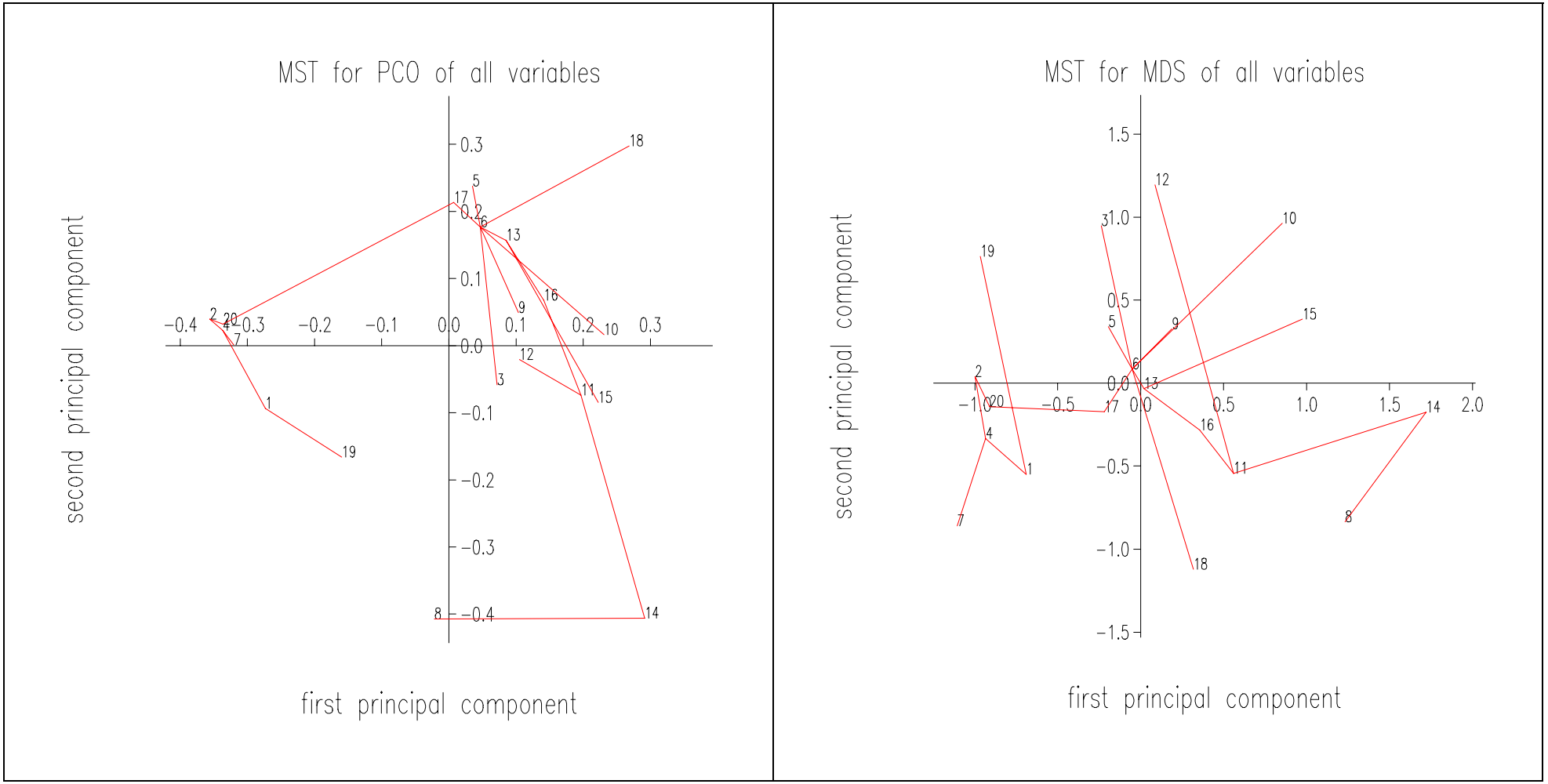


Abbildung A41: Konfigurationen der Betriebe nach Hauptkoordinatenanalyse (PCO) und mehrdimensionaler ordinaler Skalierung (MDS) der Kulturmaßnahmen in zwei Dimensionen mit überlagerten Multiple Spanning Trees

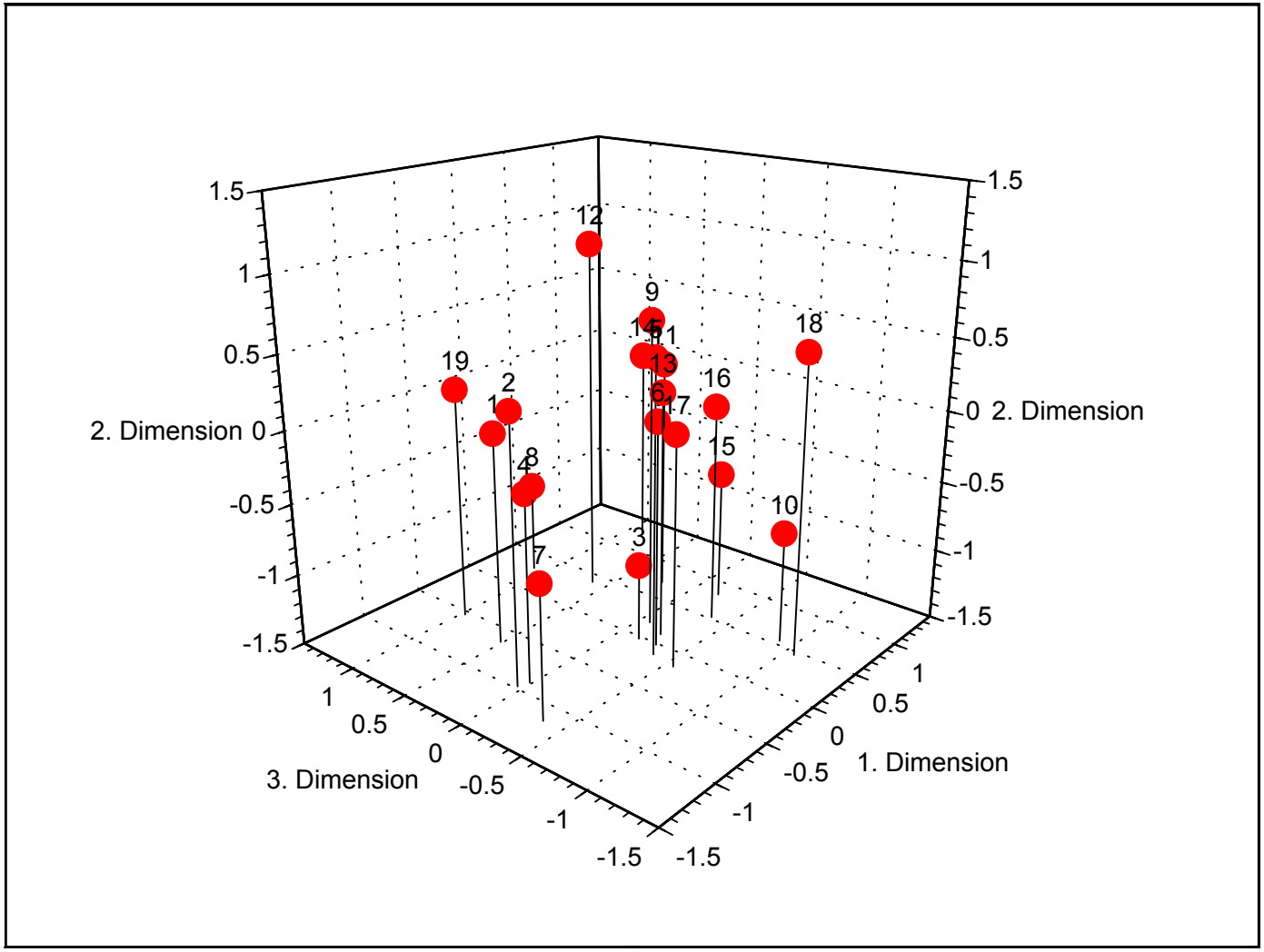


Abbildung A42: Darstellung der ersten drei Dimensionen der ordinalen mehrdimensionalen Skalierung der Kulturmaßnahmen

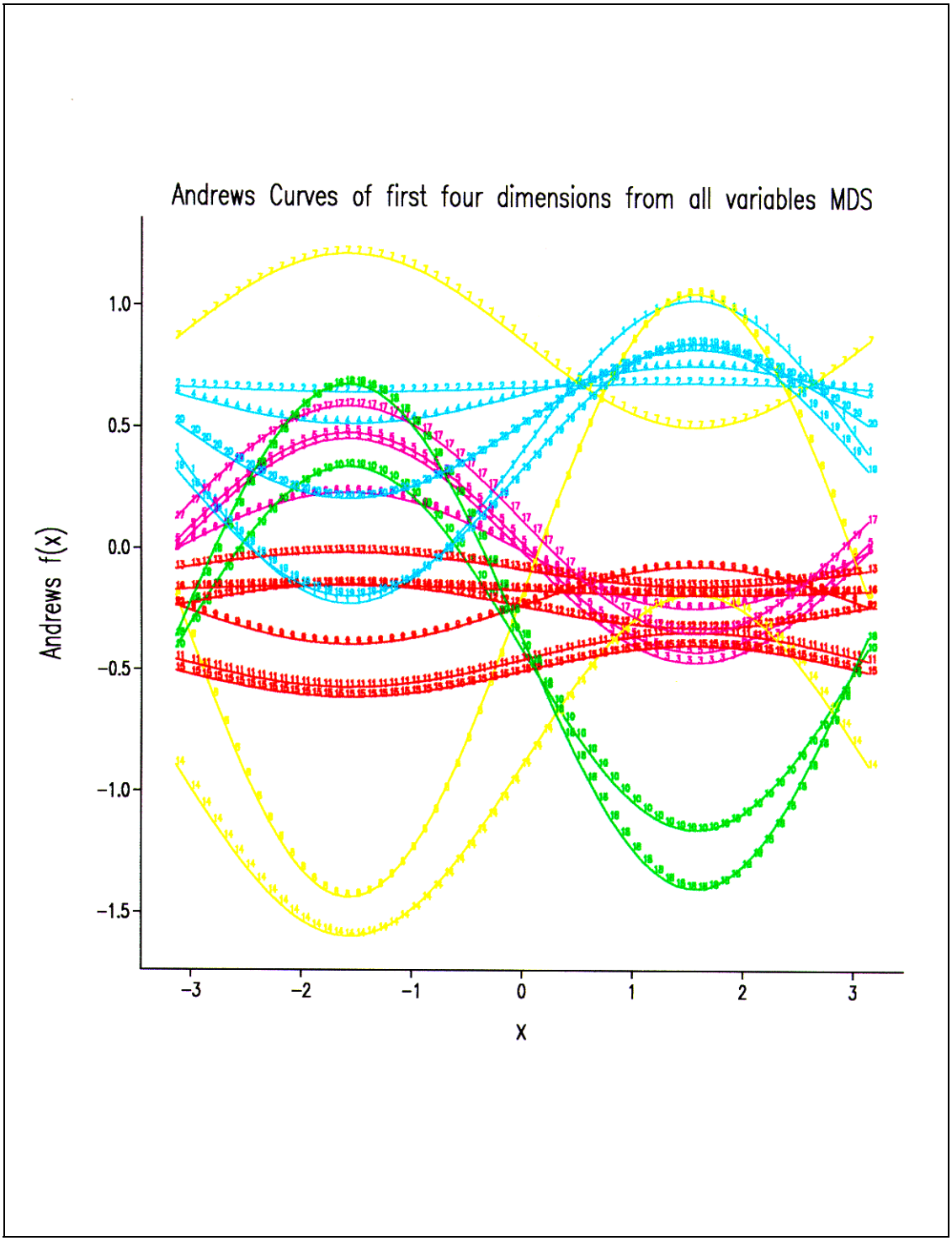


Abbildung A43: Andrews Kurven der ersten vier Dimensionen der ordinalen mehrdimensionalen Skalierung aller Variablen des Kultumaßnahmen Datensets

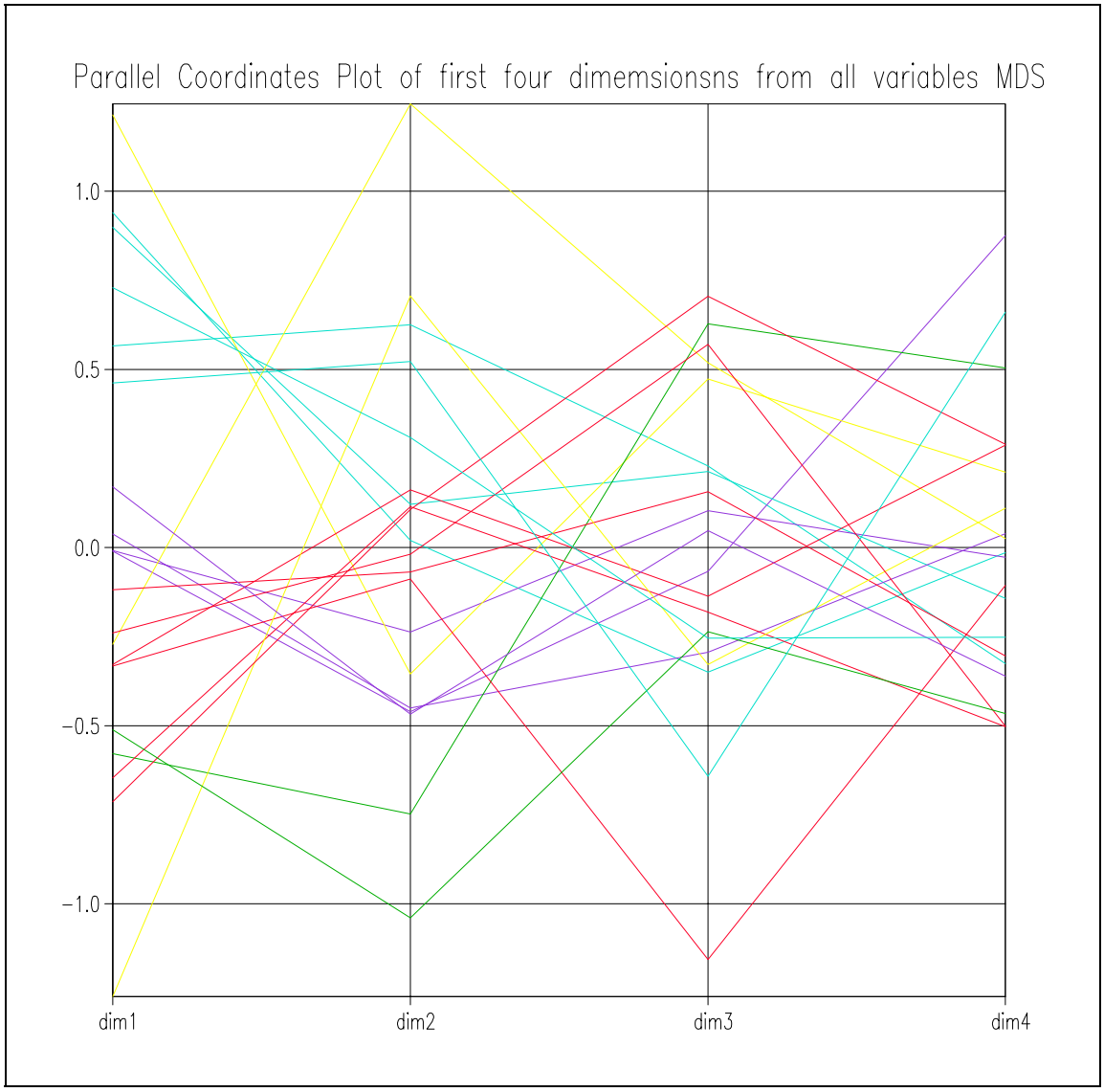


Abbildung A44: Parallelkoordinatenplot der ersten vier Dimensionen der ordinalen mehrdimensionalen Skalierung aller Variablen des Kulturmaßnahmen Datensets (farbliche Hervorhebung der aus dem Andrews-Plot abgeleiteten Gruppierung)

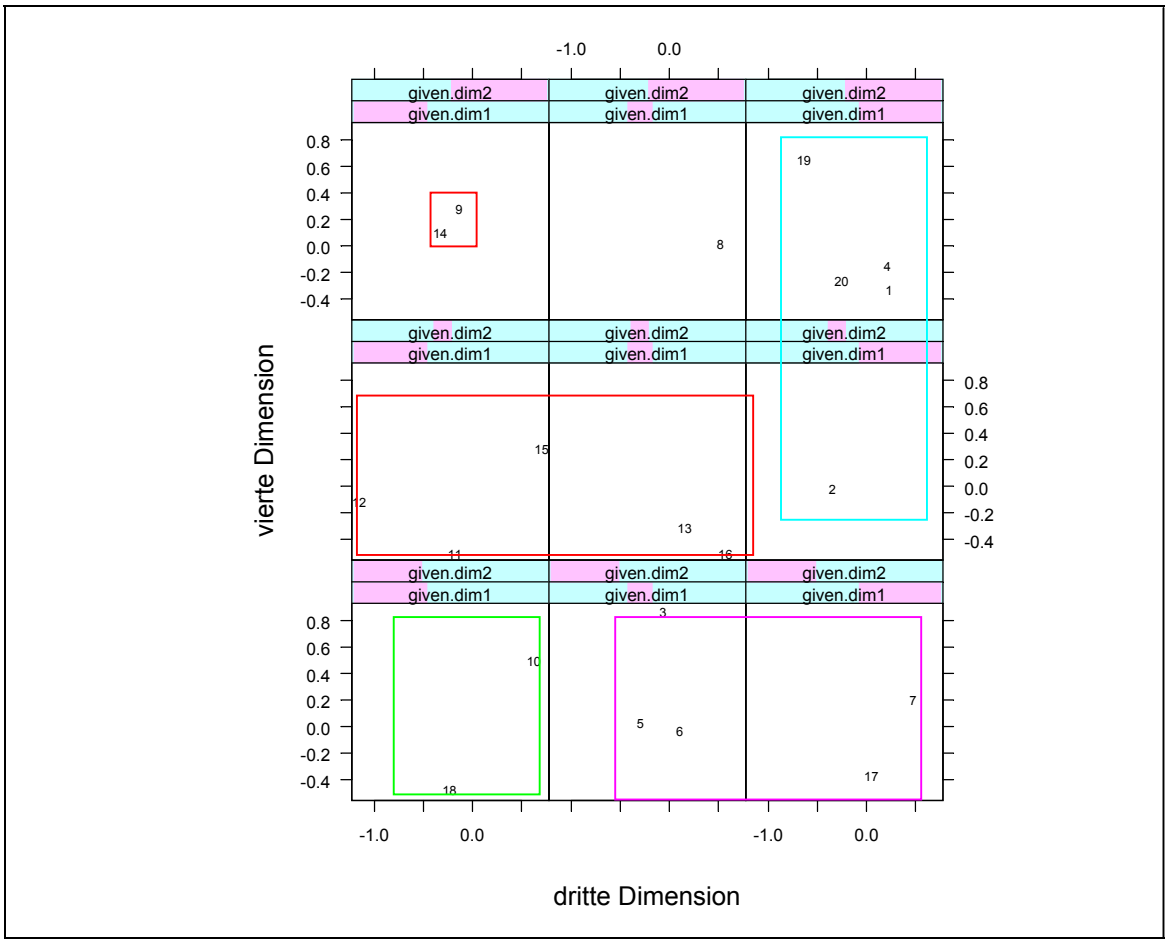


Abbildung A45: Trellis-Display der dritten und vierten Dimension, konditioniert durch die erste und zweite Dimension (given.dim1 beziehungsweise given.dim2)

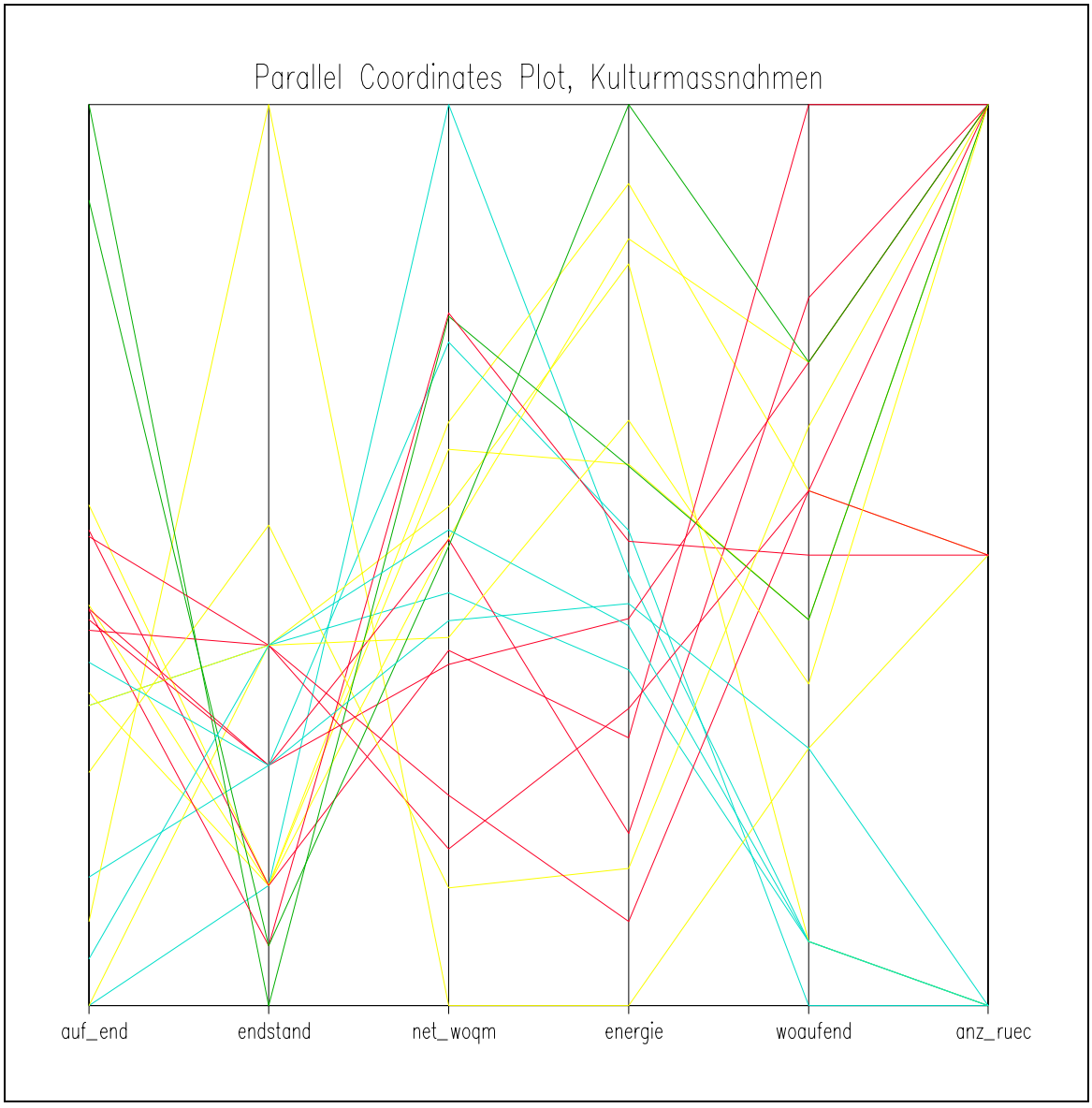


Abbildung A46: Parallelkoordinatenplot ausgewählter Variablen des Datensets 3 (Kulturmaßnahmen) mit farblicher Hervorhebung der aus dem Andrews-Plot abgeleiteten Gruppierung

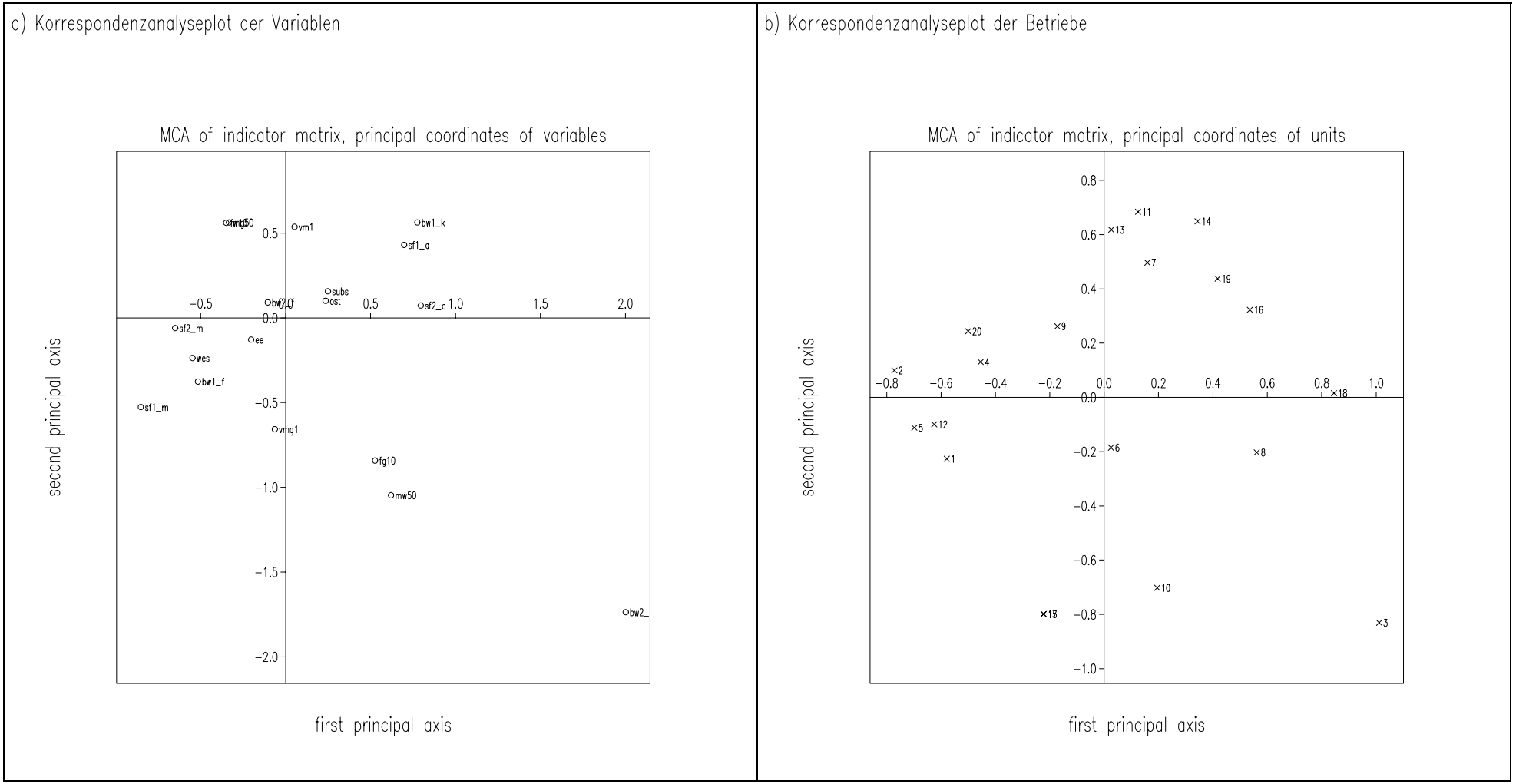


Abbildung A47: Korrespondenzanalyseplot der Variablen(a)) und der Betriebe (b)) im Variablenset 4; durch die erste Dimension erklärte Varianz 25,5%, durch die zweite Dimension erklärte Varianz 22,9%

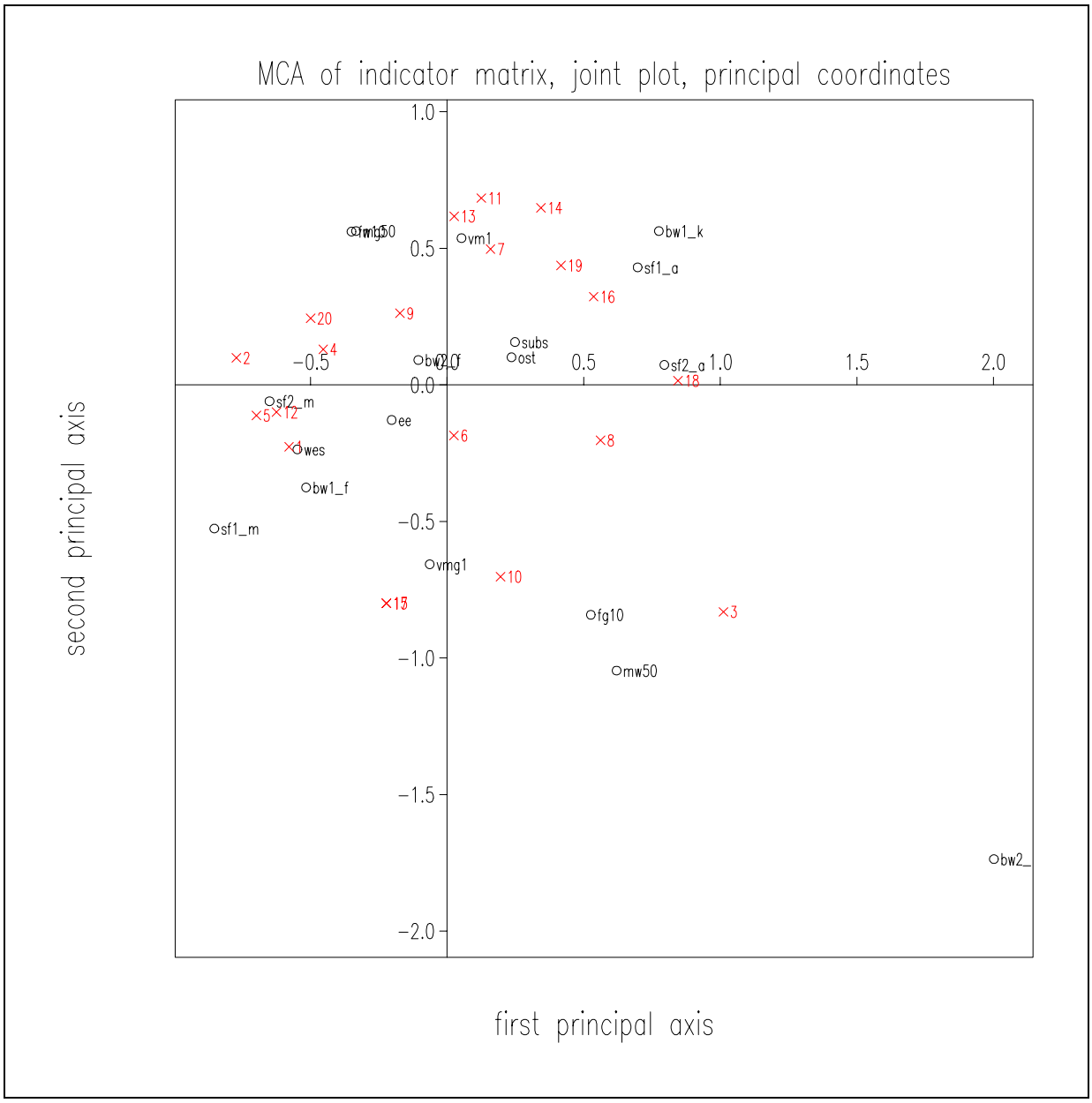


Abbildung A48: Gemeinsamer Korrespondenzanalyseplot der Variablen und der Betriebe in Normalkoordinaten; durch die erste Dimension erklärte Varianz 25,5%, durch die zweite Dimension erklärte Varianz 22,9%

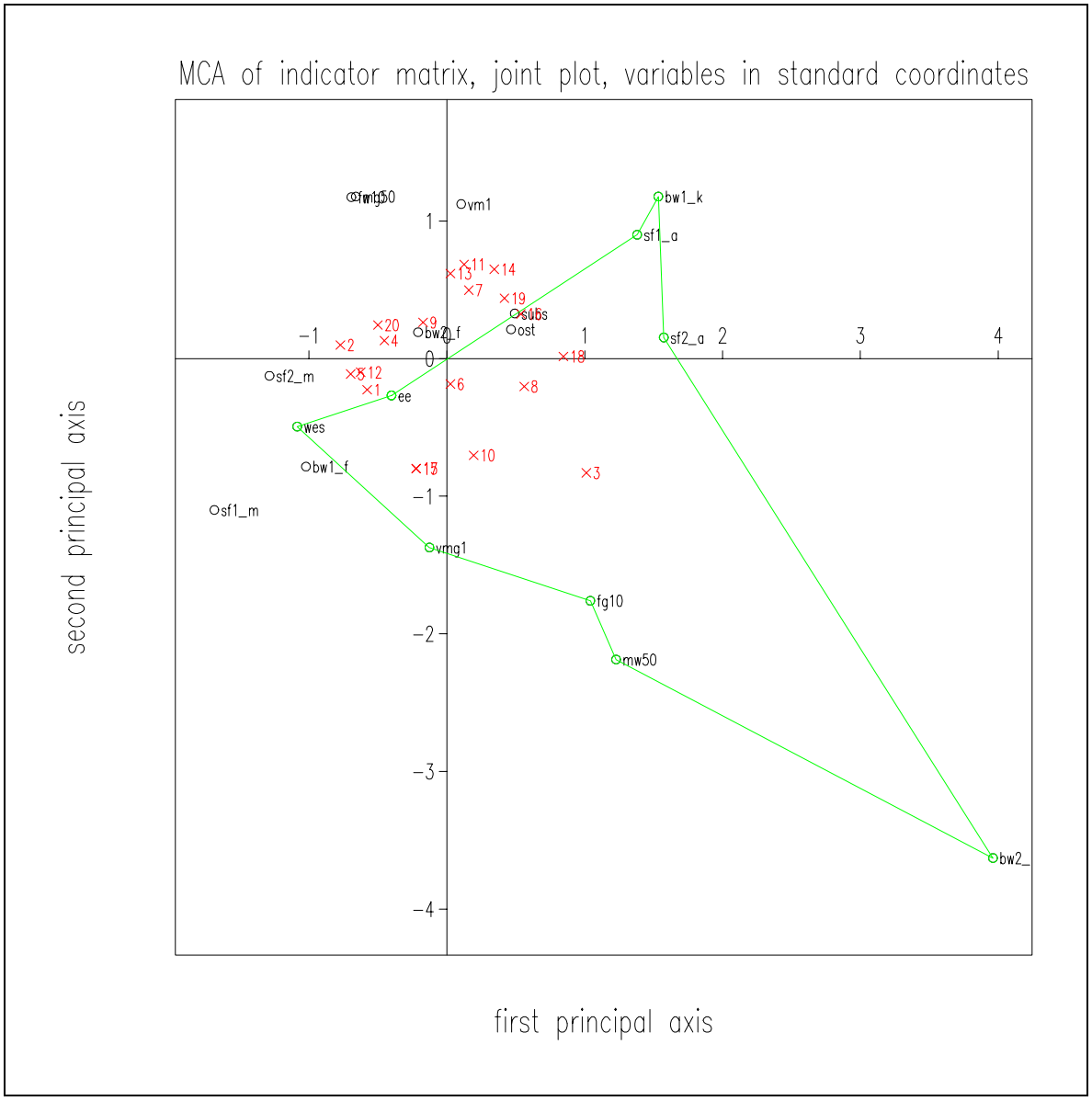


Abbildung A49: Gemeinsamer Korrespondenzanalyseplot der Variablen in Standard- und der Betriebe in Normalkoordinaten mit Interpolationsregion für Betrieb 3; erklärte Varianz durch die erste Dimension 25,5%, durch die zweite Dimension 22,9%

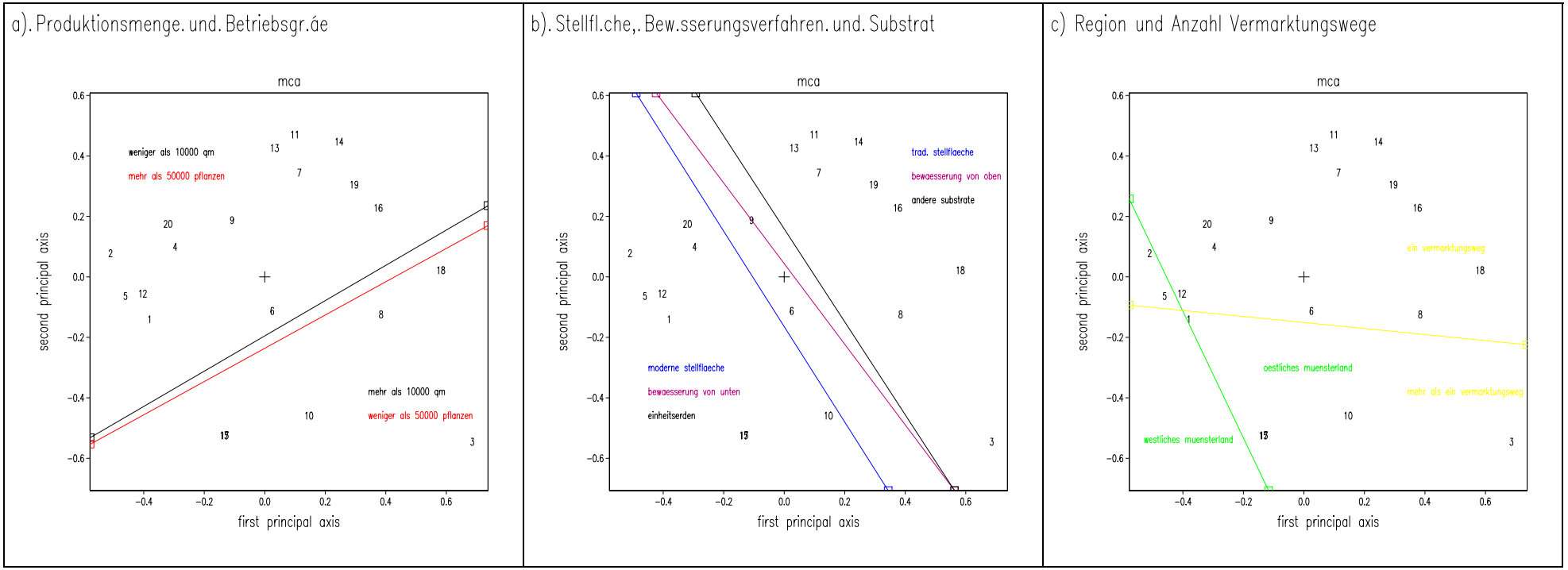


Abbildung A50: Prediktionsregionen der Korrespondenzanalyse in getrennten Plots für einzelne Variablen basierend auf der Chi-Quadrat-Distanz (mca); durch die erste Dimension erklärte Varianz 25,5%, durch die zweite Dimension erklärte Varianz 22,9%

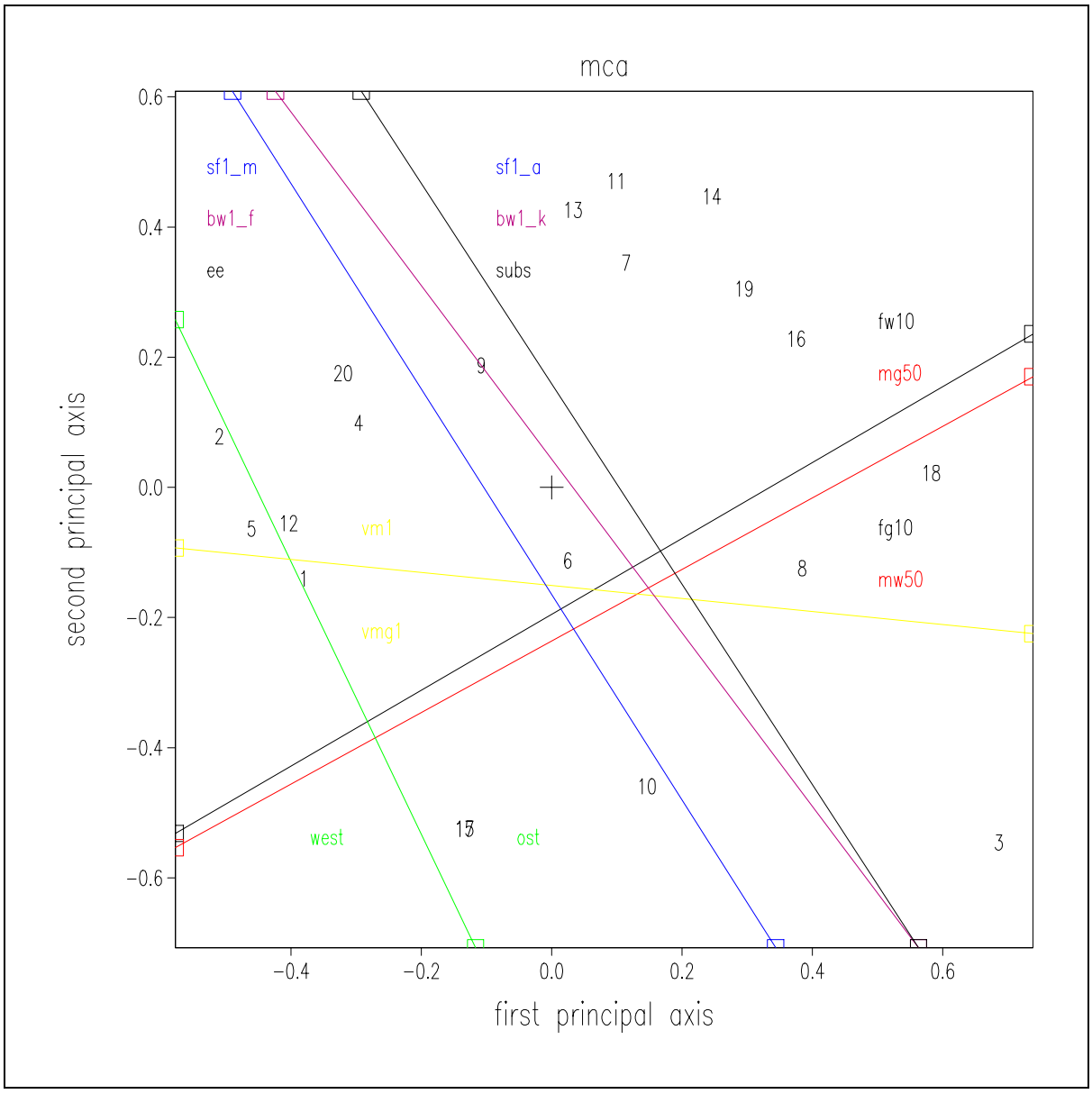


Abbildung A51: Prediktionsregionen der Korrespondenzanalyse basierend auf der Chi-Quadrat-Distanz (mca); durch die erste Dimension erklärte Varianz 25,5%, durch die zweite Dimension erklärte Varianz 22,9%

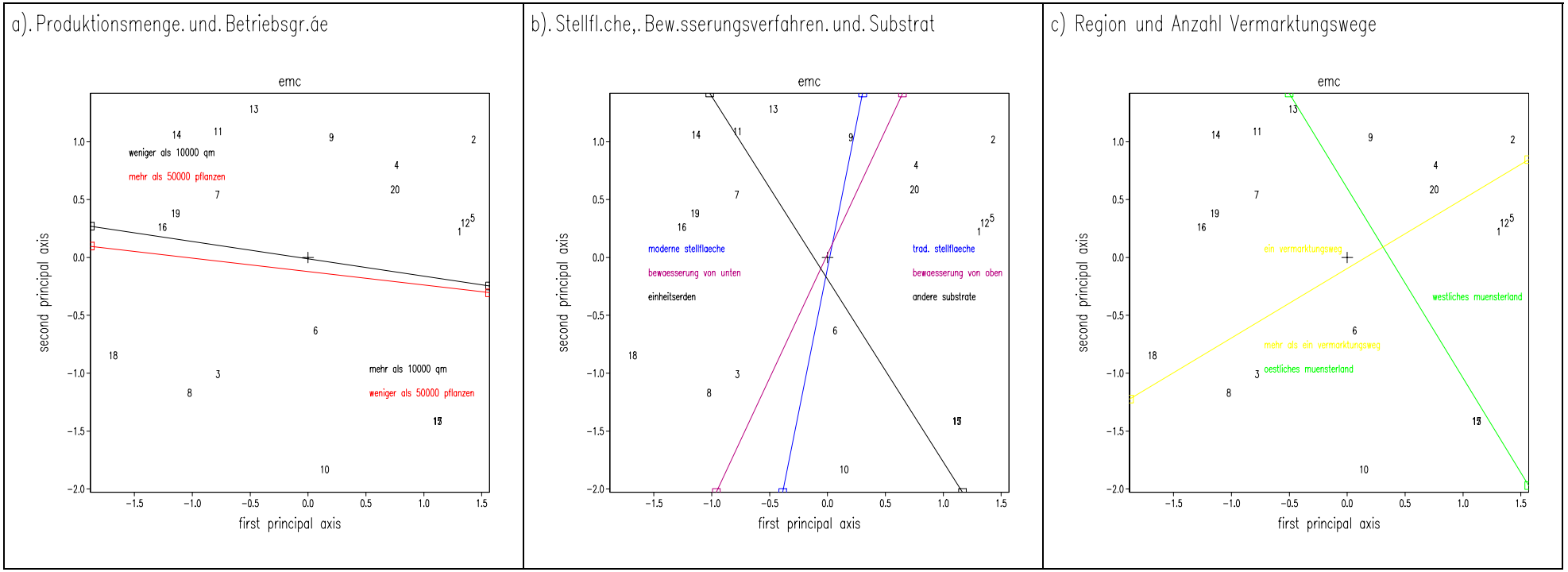


Abbildung A52: Prediktionsregionen der Korrespondenzanalyse in getrennten Plots für einzelne Variablen basierend auf dem extended matching coefficient (emc); durch die erste Dimension erklärte Varianz 27,6%, durch die zweite Dimension erklärte Varianz 23,9%

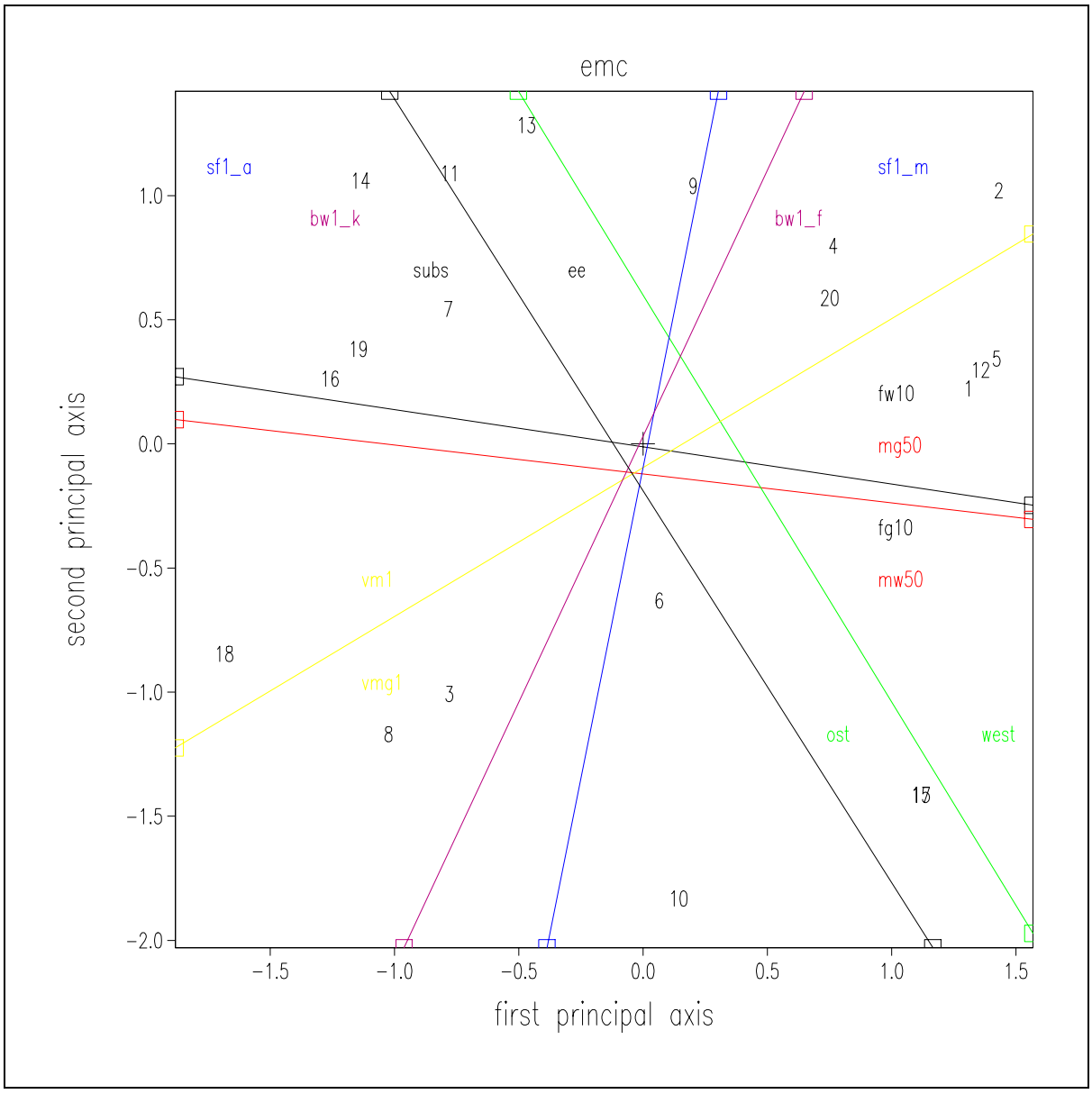


Abbildung A53: Prediktionsregionen der Korrespondenzanalyse basierend auf dem extended matching coefficient (emc); durch die erste Dimension erklärte Varianz 27,6%, durch die zweite Dimension erklärte Varianz 23,9%

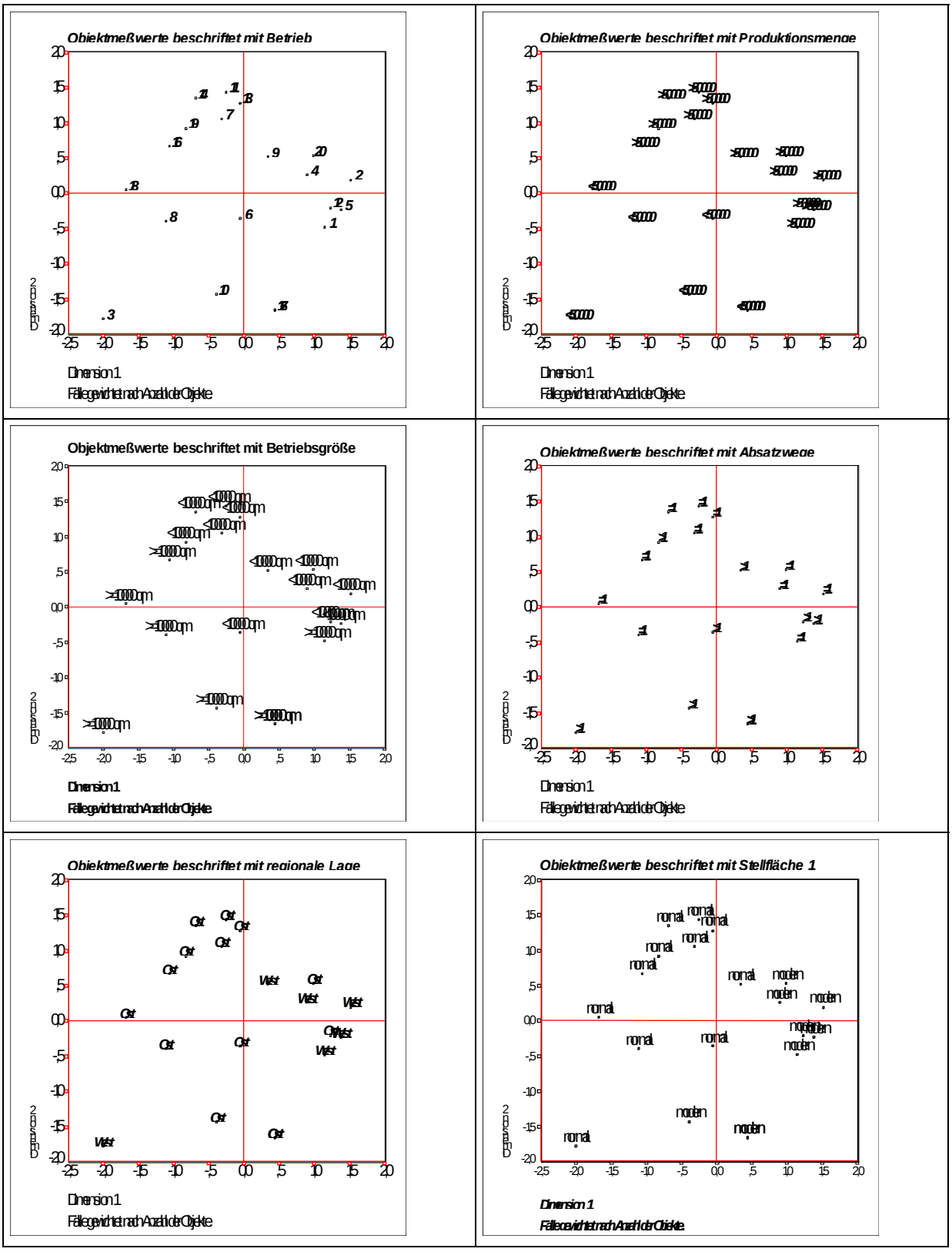


Abbildung A54: Darstellung der Korrespondenzanalyse-Konfiguration durch beschriftete Objektmeßwerte-Plots

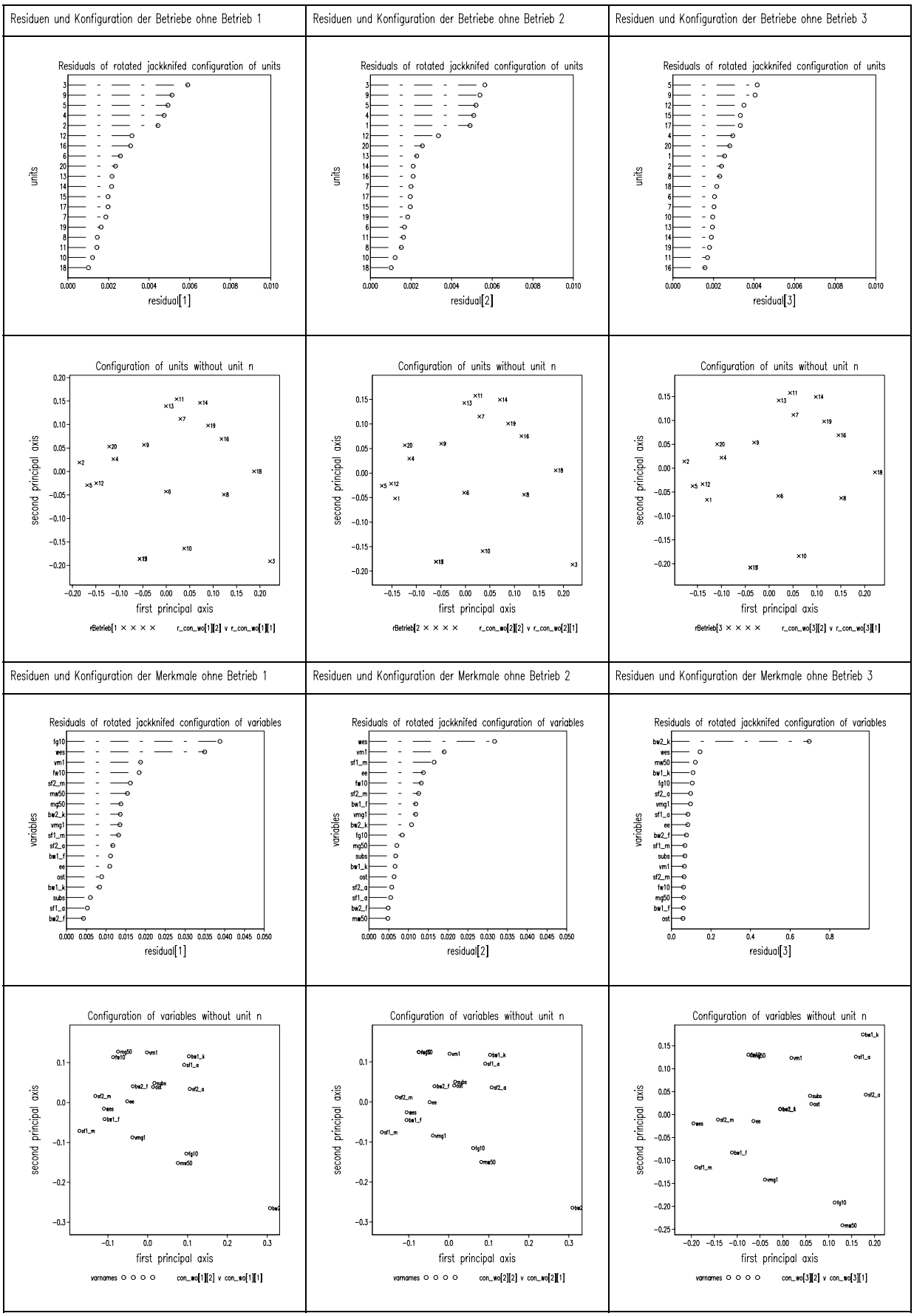


Abbildung A55: Residuen zur Konsenz-Konfiguration der Betriebe und der der Merkmale ohne die Objekte 1, 2 und 3; Konfigurationen der Betriebe und der Merkmale der Korrespondenzanalyse der Strukturmerkmale ohne die Objekte 1, 2 und 3

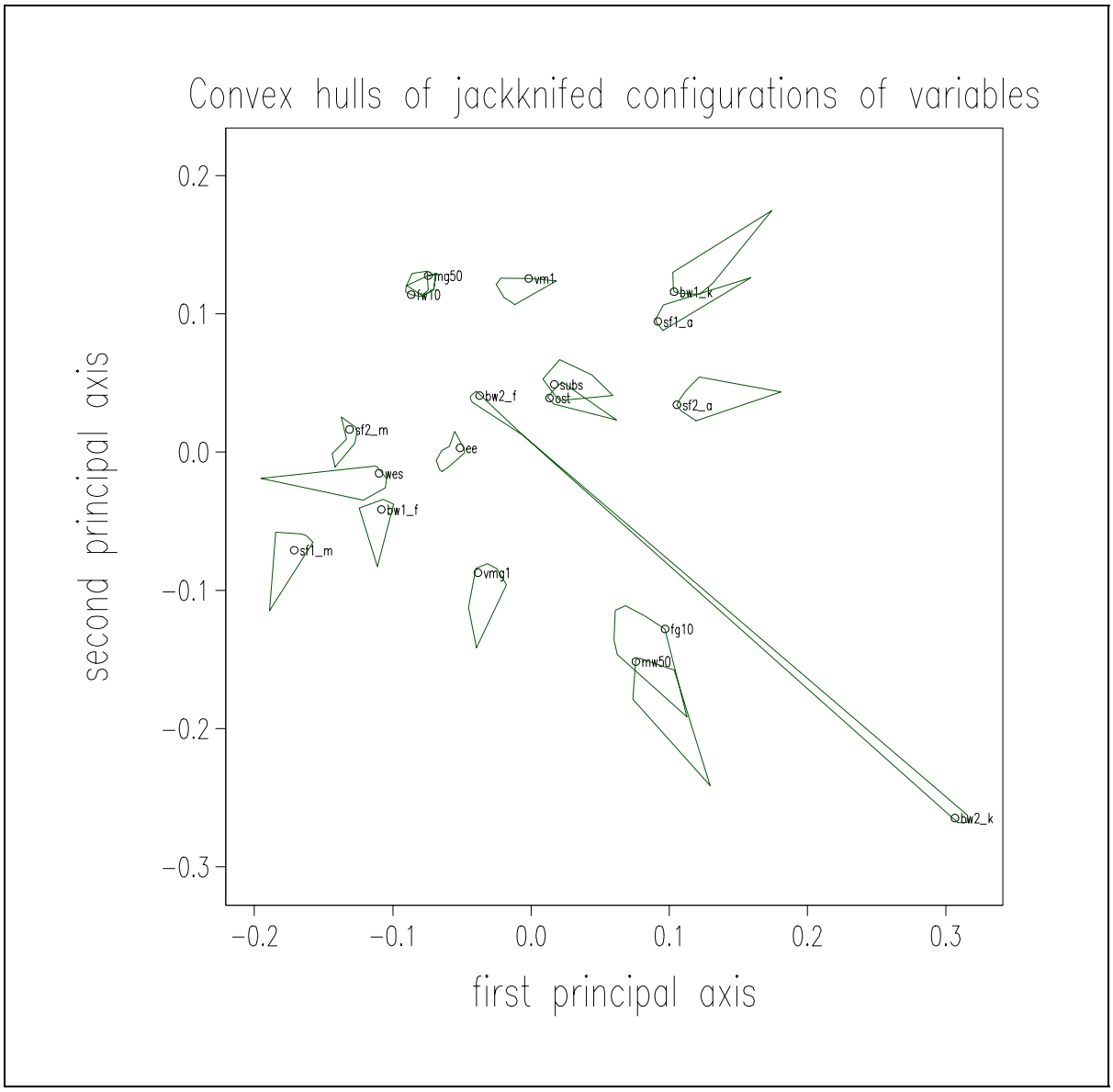


Abbildung A56: Beurteilung der Stabilität der Positionen der Variablen in der Korrespondenzanalyse der Strukturmerkmale durch konvexe Hüllen

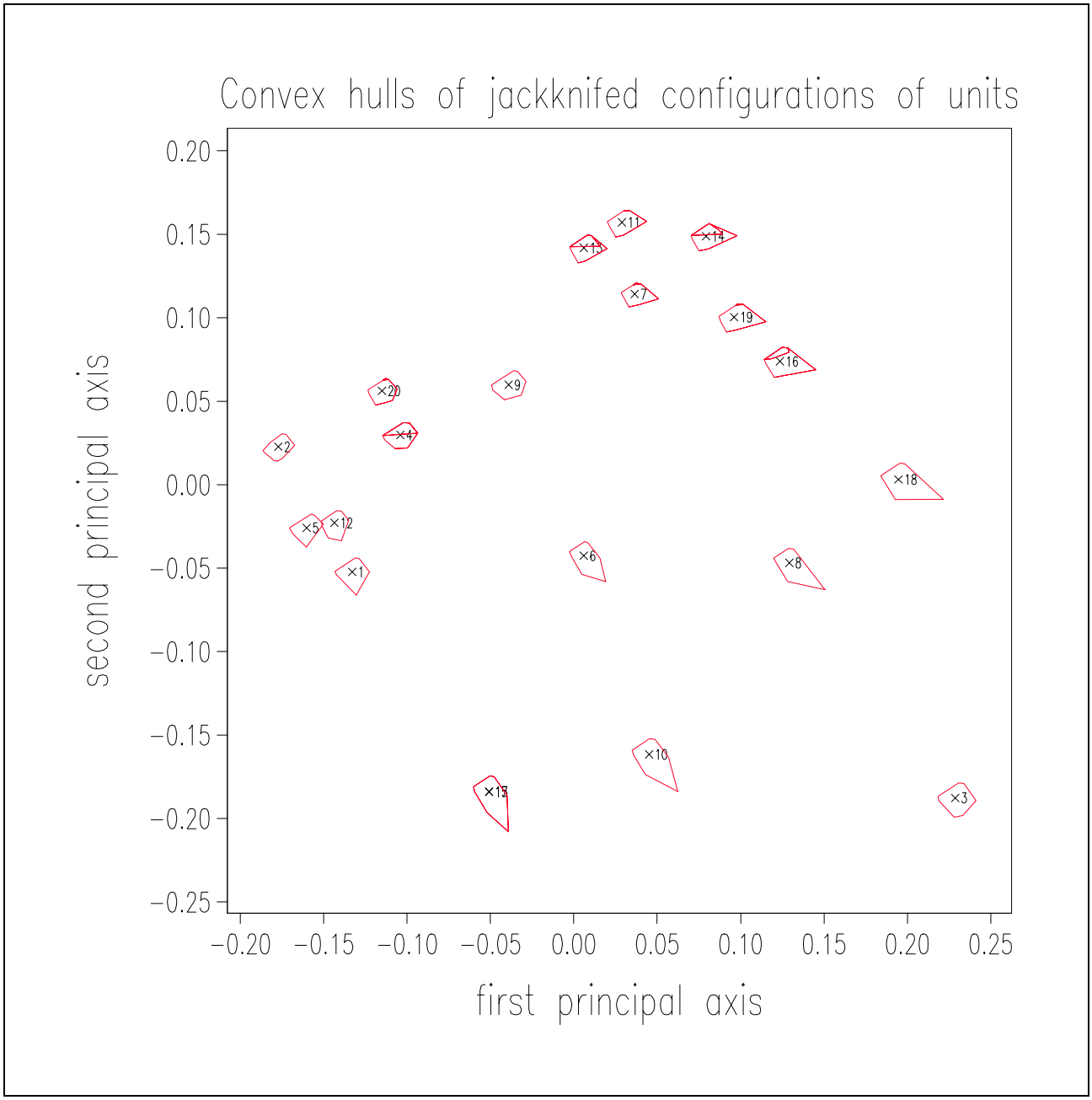


Abbildung A57: Beurteilung der Stabilität der Positionen der Objekte in der Korrespondenzanalyse der Strukturmerkmale durch konvexe Hüllen

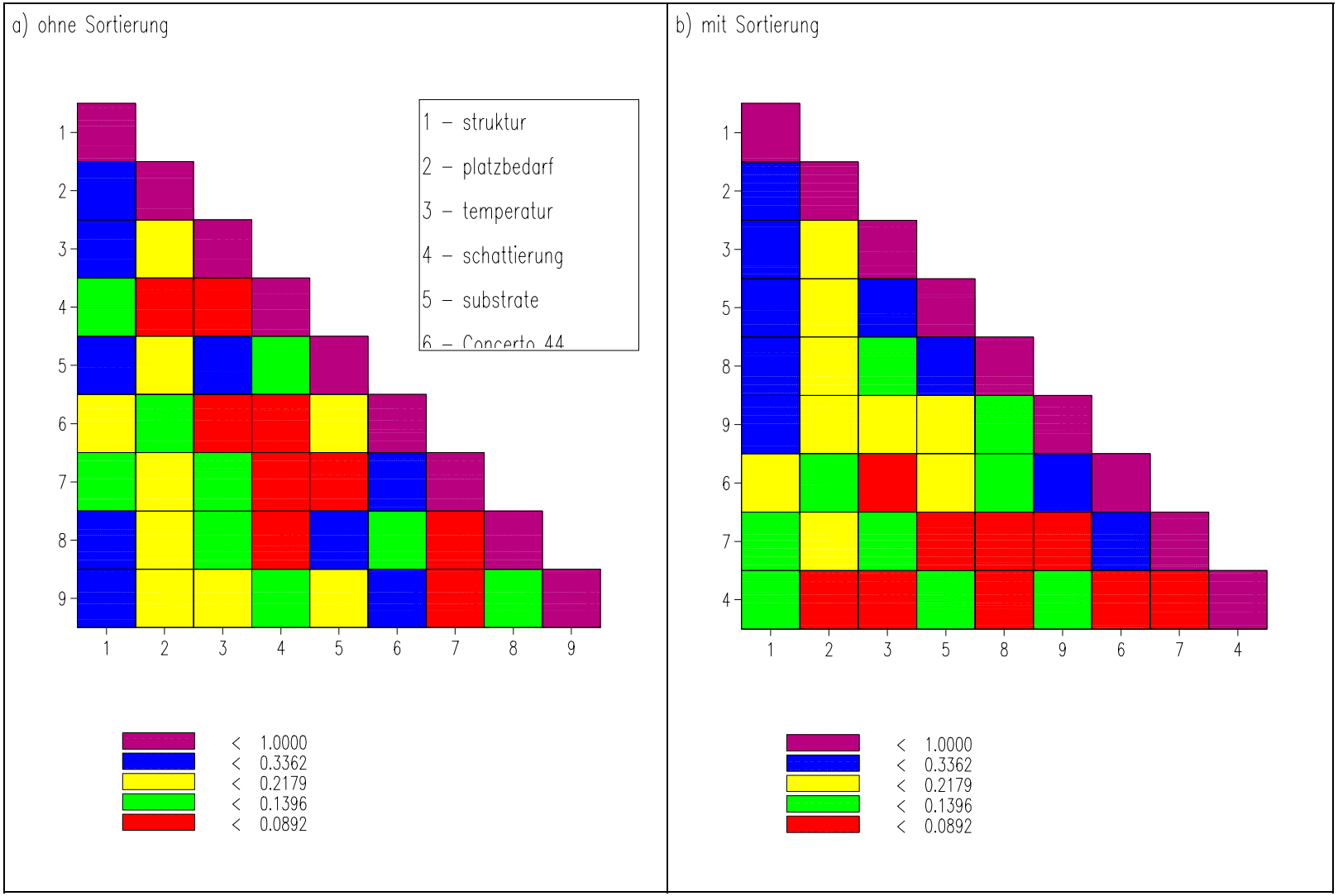


Abbildung A58: Dshade-Diagramme der Proximitätsmatrix der paarweisen Residuen der multiplen Prokrustes-Rotation aller Variablensets

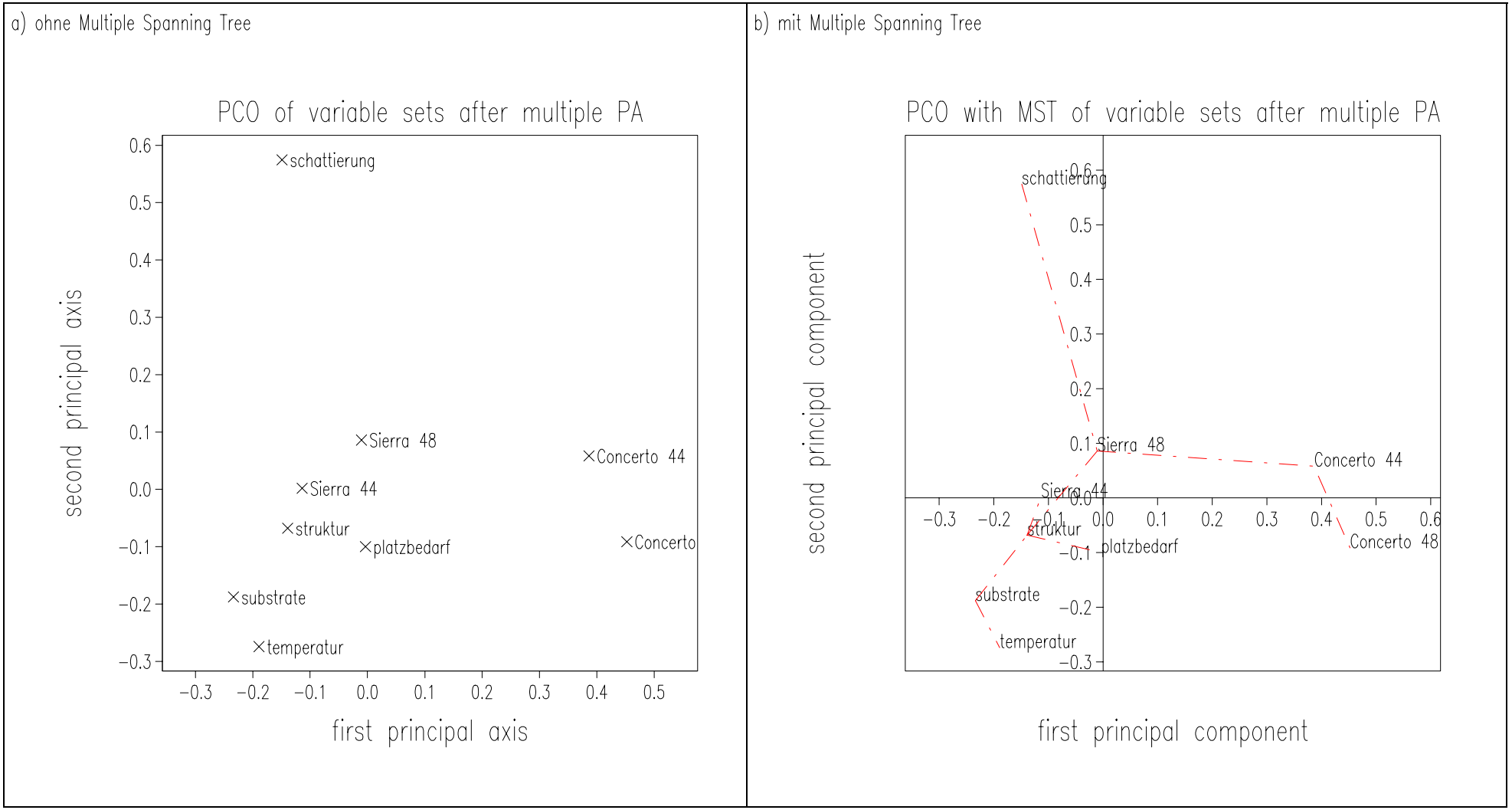


Abbildung A59: Hauptkoordinatenanalyse der Proximitätsmatrix der paarweisen Residuen der multiplen Prokrustes-Rotation aller Variablensets; Anteil erklärter Varianz durch die erste Dimension 16,4%, durch die zweite Dimension 15,5%

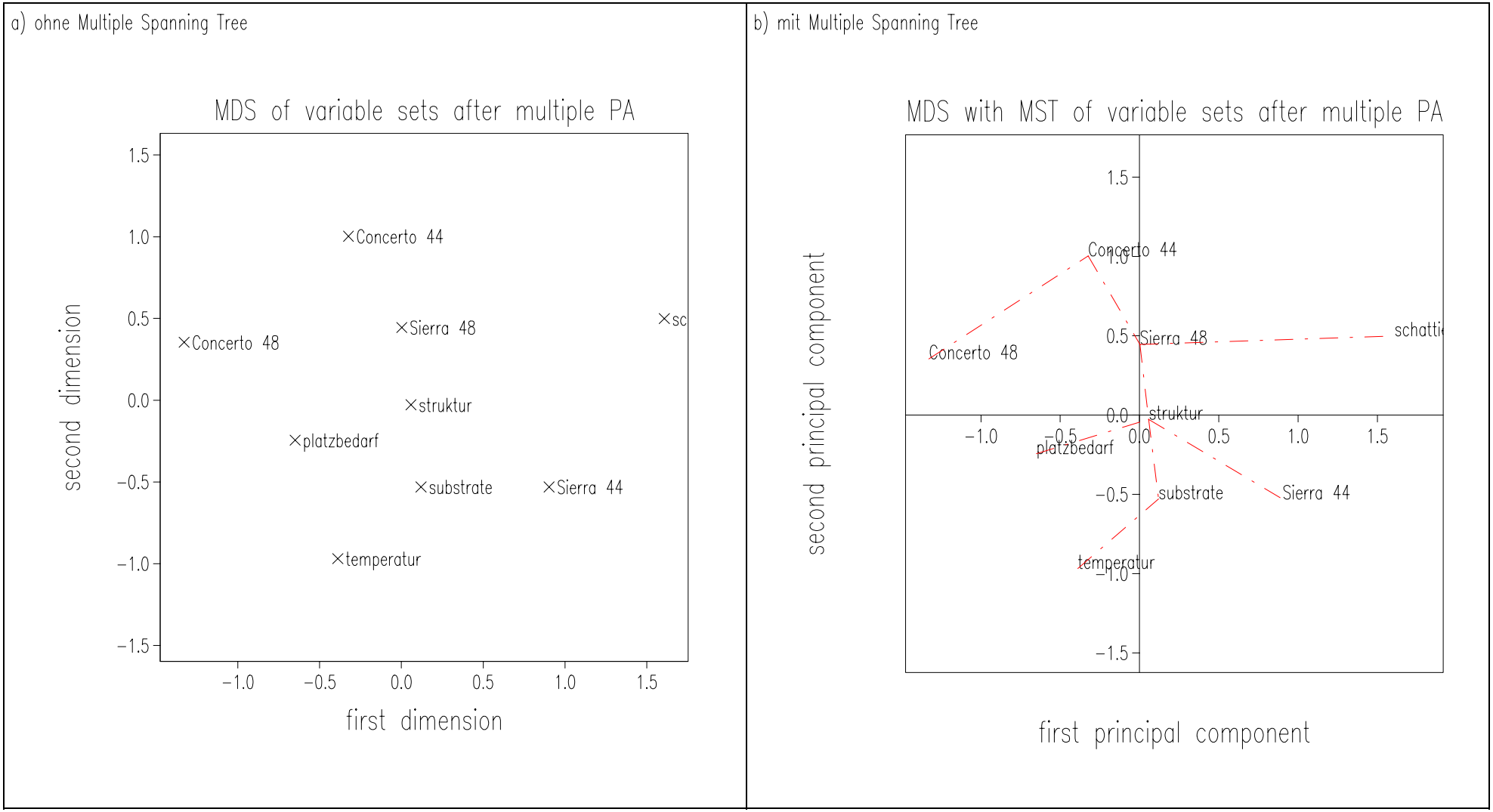


Abbildung A60: Ordinale mehrdimensionale Skalierung der Proximitätsmatrix der paarweisen Residuen der multiplen Prokrustes-Rotation aller Variablensets; Stress in zwei Dimensionen 0,1220

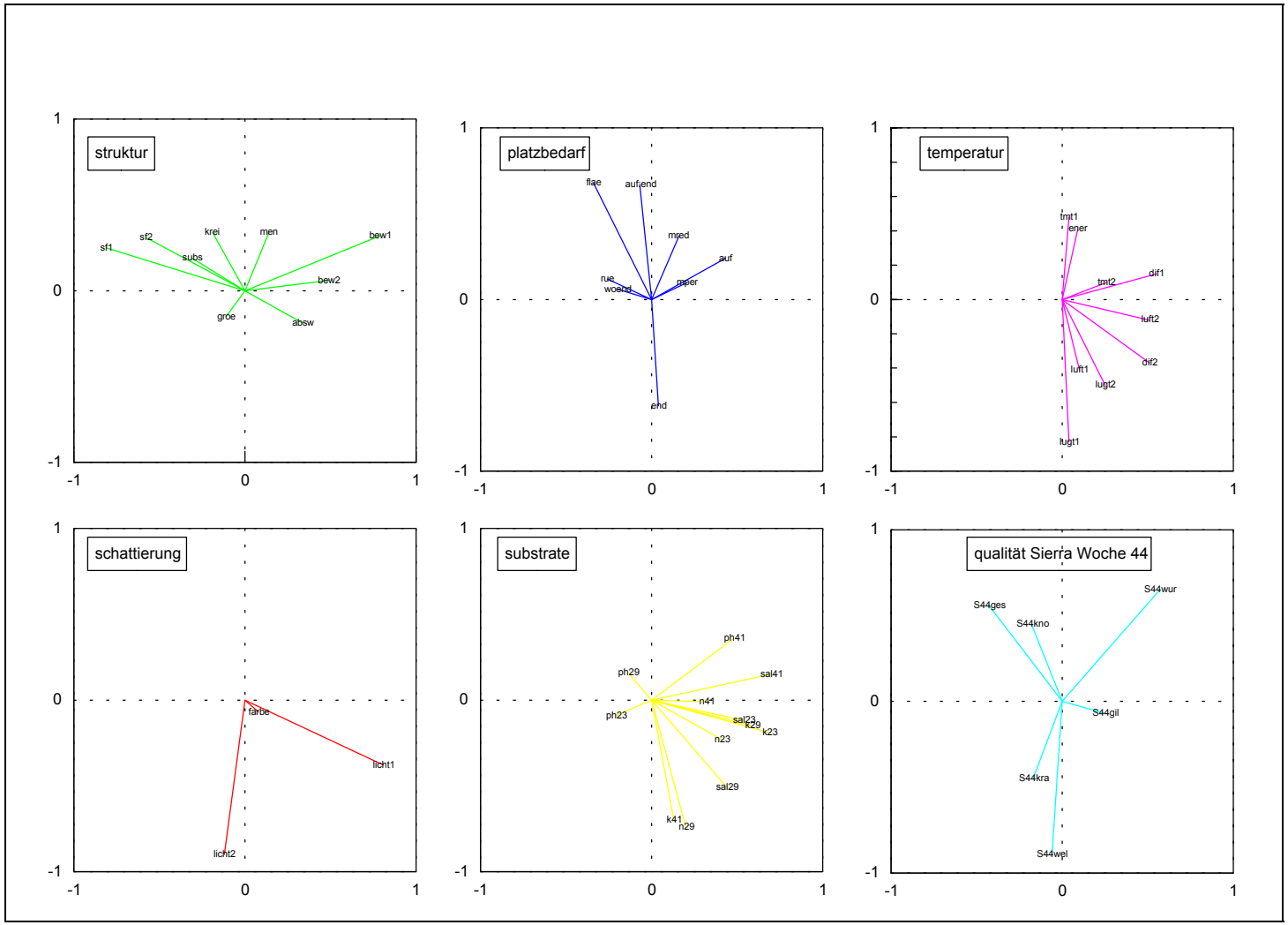


Abbildung A61: Komponentenladungen der generalisierten kanonischen Analyse, ‘Sierra’ Woche 44 im Datenset 6; mittlerer Loss 0,105

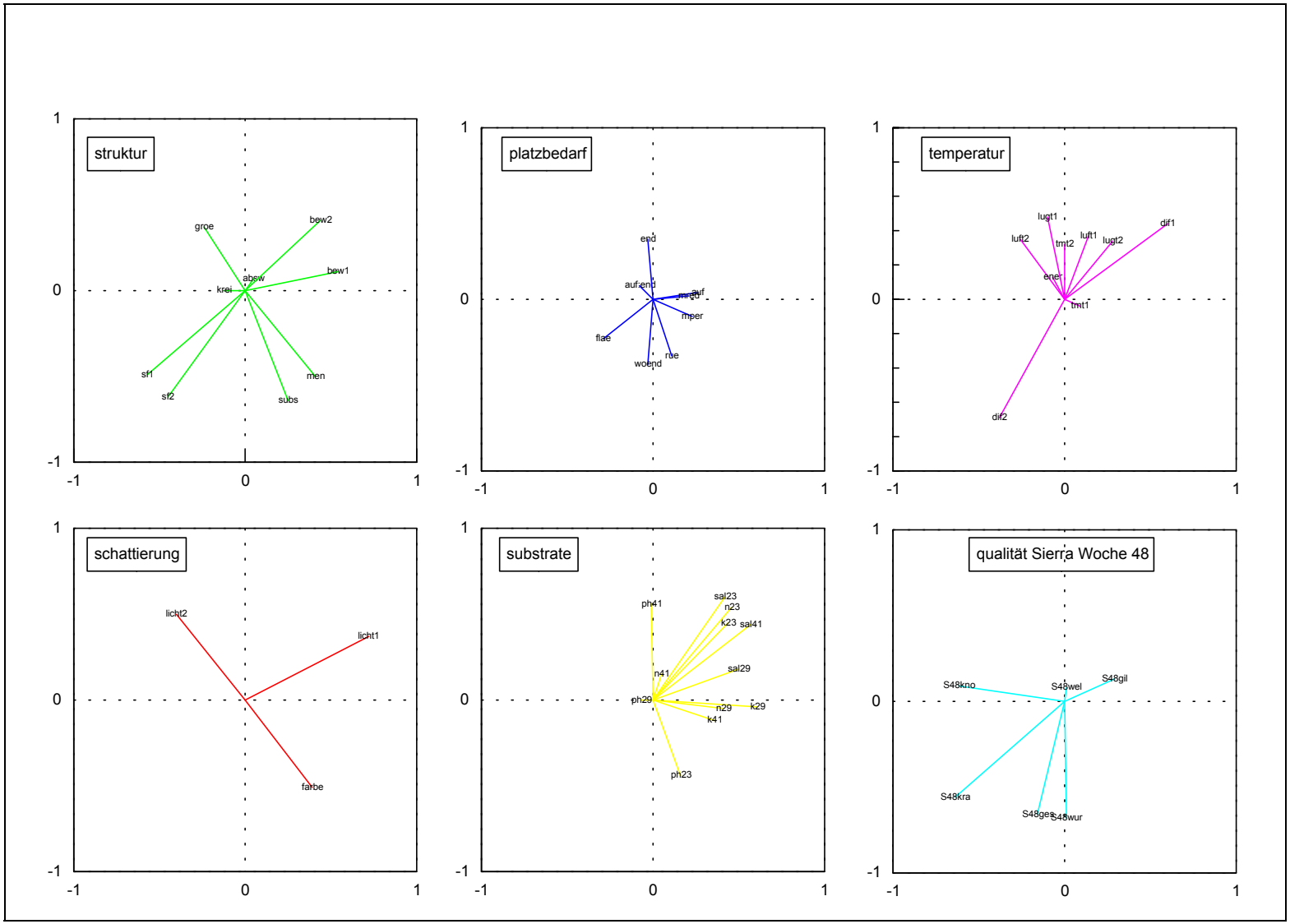


Abbildung A62: Komponentenladungen generalisierter kanonischen Analyse, ‘Sierra’ Woche 48 im Datenset 6; mittlerer Loss 0,044

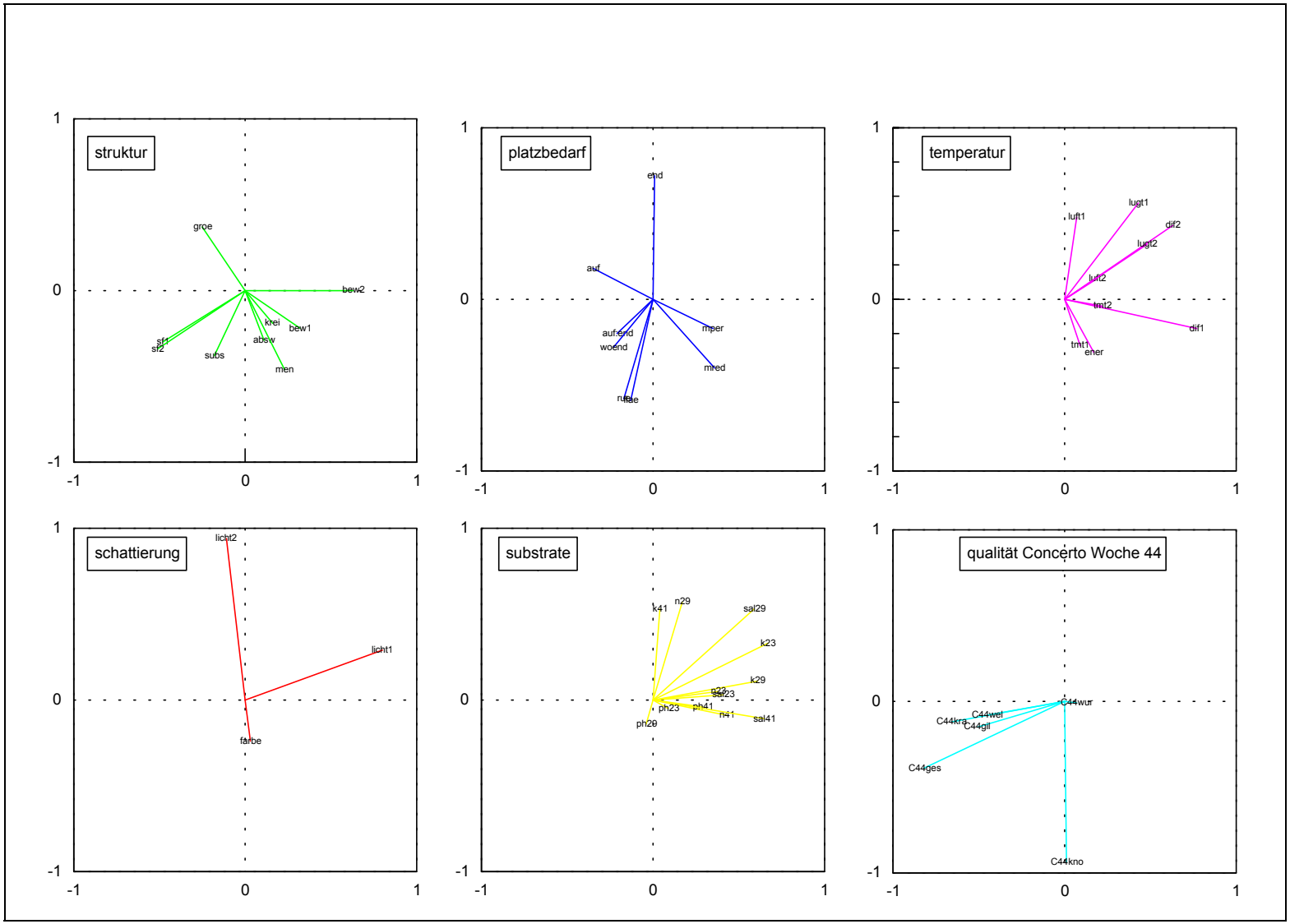


Abbildung A63: Komponentenladungen generalisierter kanonischen Analyse, ‘Concerto’ Woche 44 im Datenset 6; mittlerer Loss 0,083

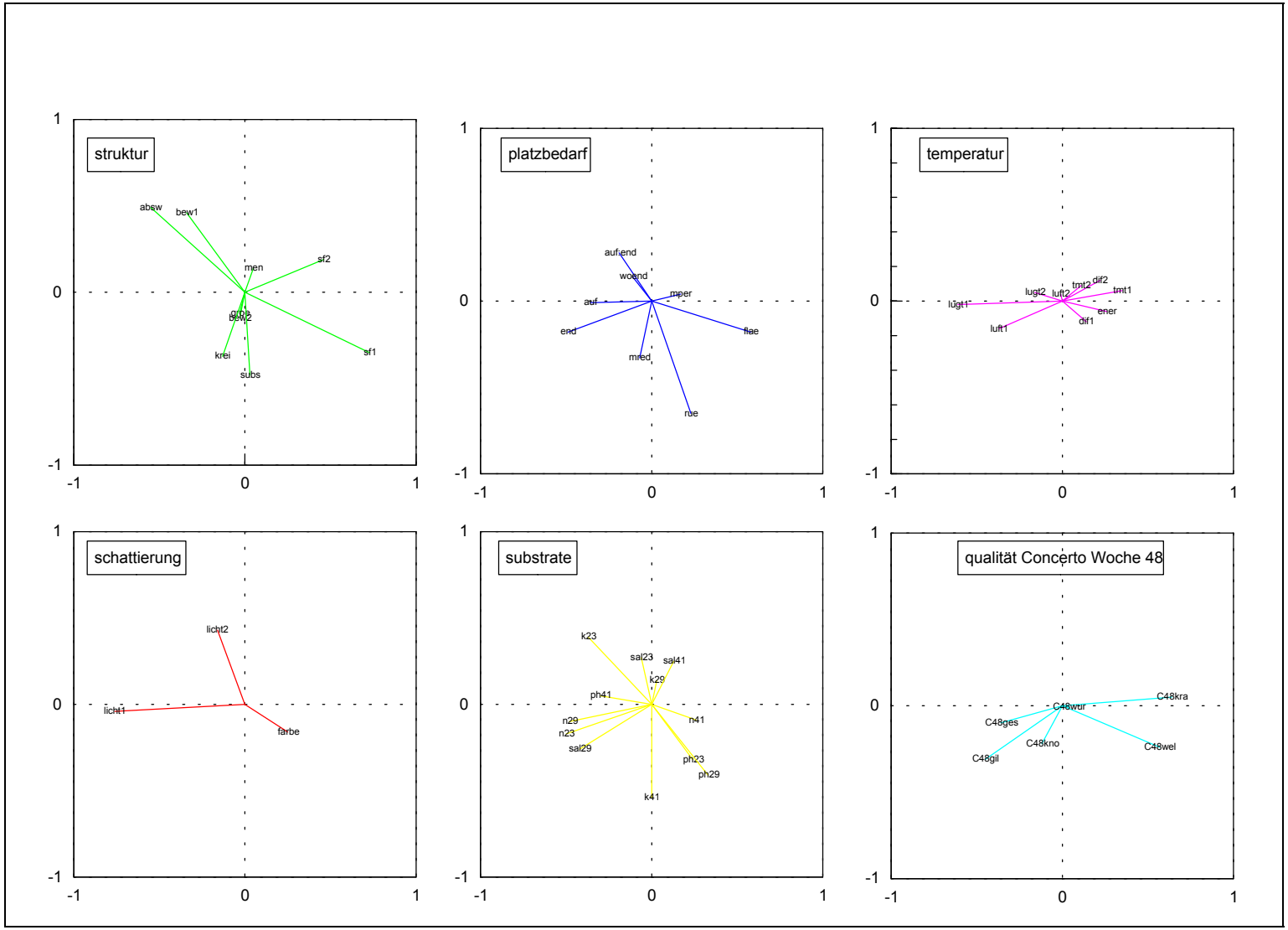


Abbildung A64: Komponentenladungen generalisierten kanonischen Analyse, ‘Concerto’ Woche 48 im Datenset 6; mittlerer Loss 0,116

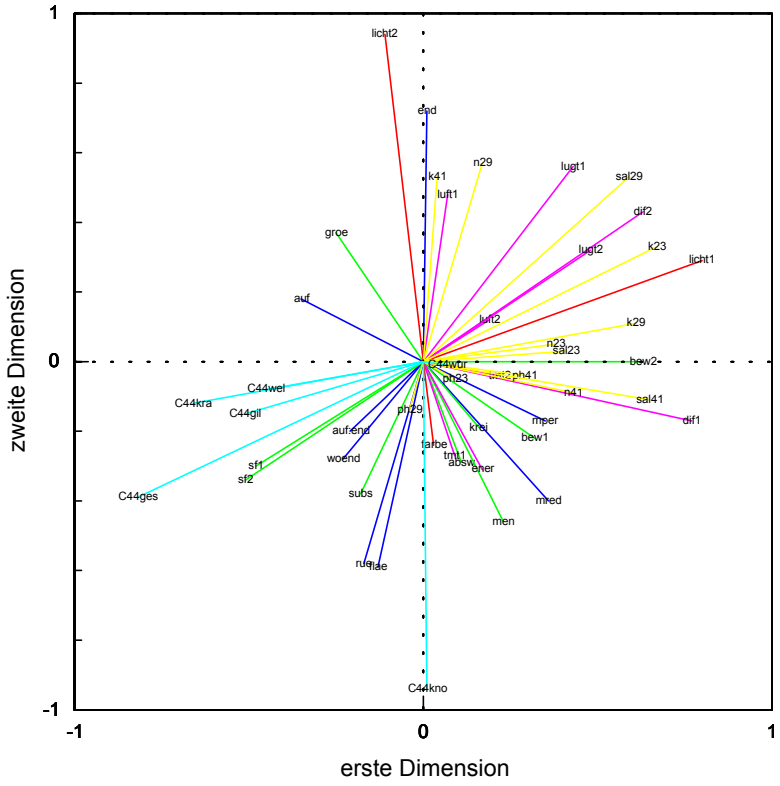


Abbildung A65: Überlagerte Komponentenladungen der generalisierten kanonischen Analyse aller Variablensets, ‘Concerto’ Woche 44 im Datenset 6; mittlerer Loss 0,083

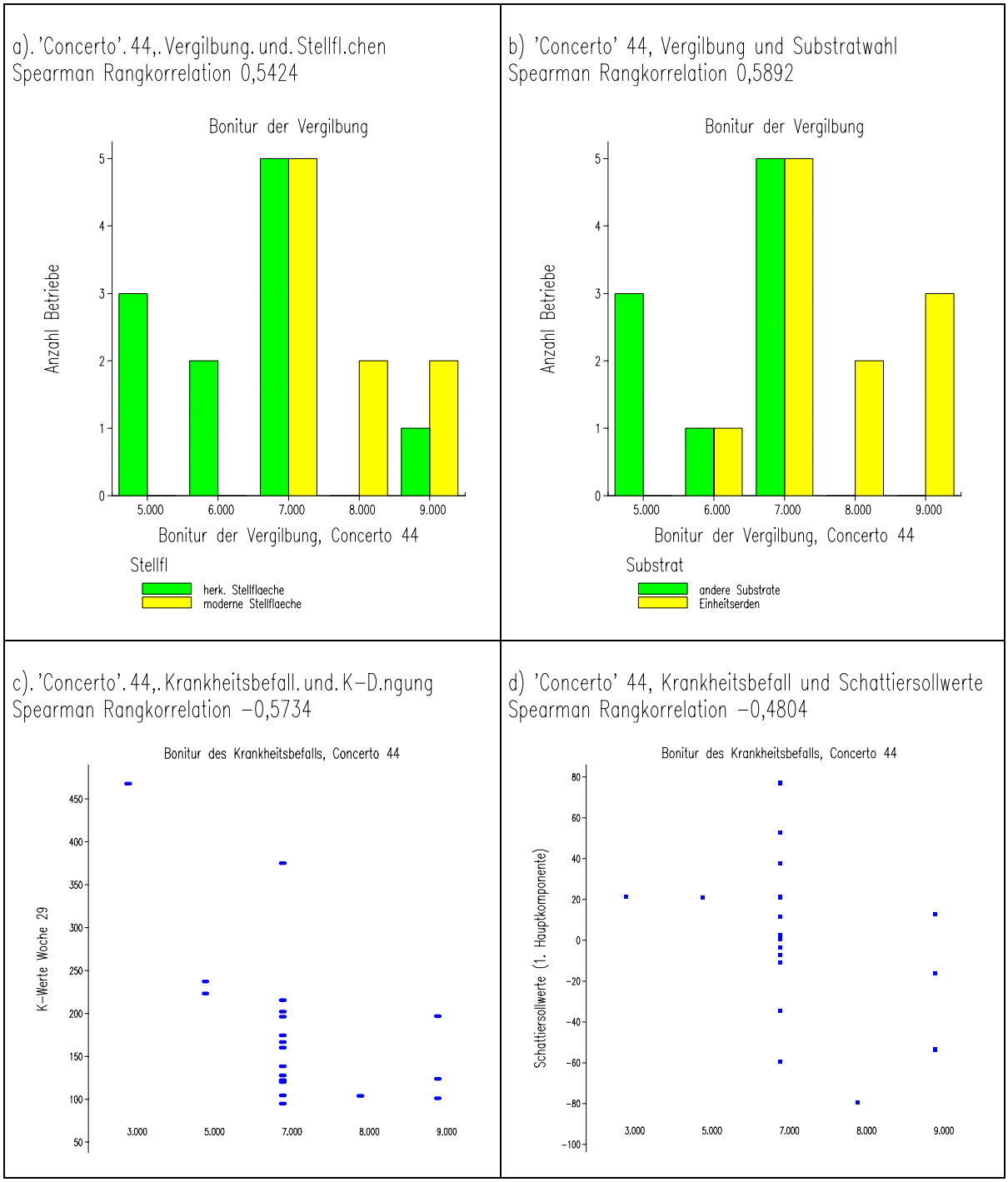


Abbildung A66: Illustration der Ergebnisse der generalisieren kanonischen Analyse nach Identifikation auffälliger Zusammenhänge bei 'Concerto' Woche 44 in Variablenset 6

Anhang Teil I B

Abbildungen zur Auswertung der Kennzahlen der Topfpflanzenbetriebe 1992 bis 1994, Kapitel 3.2

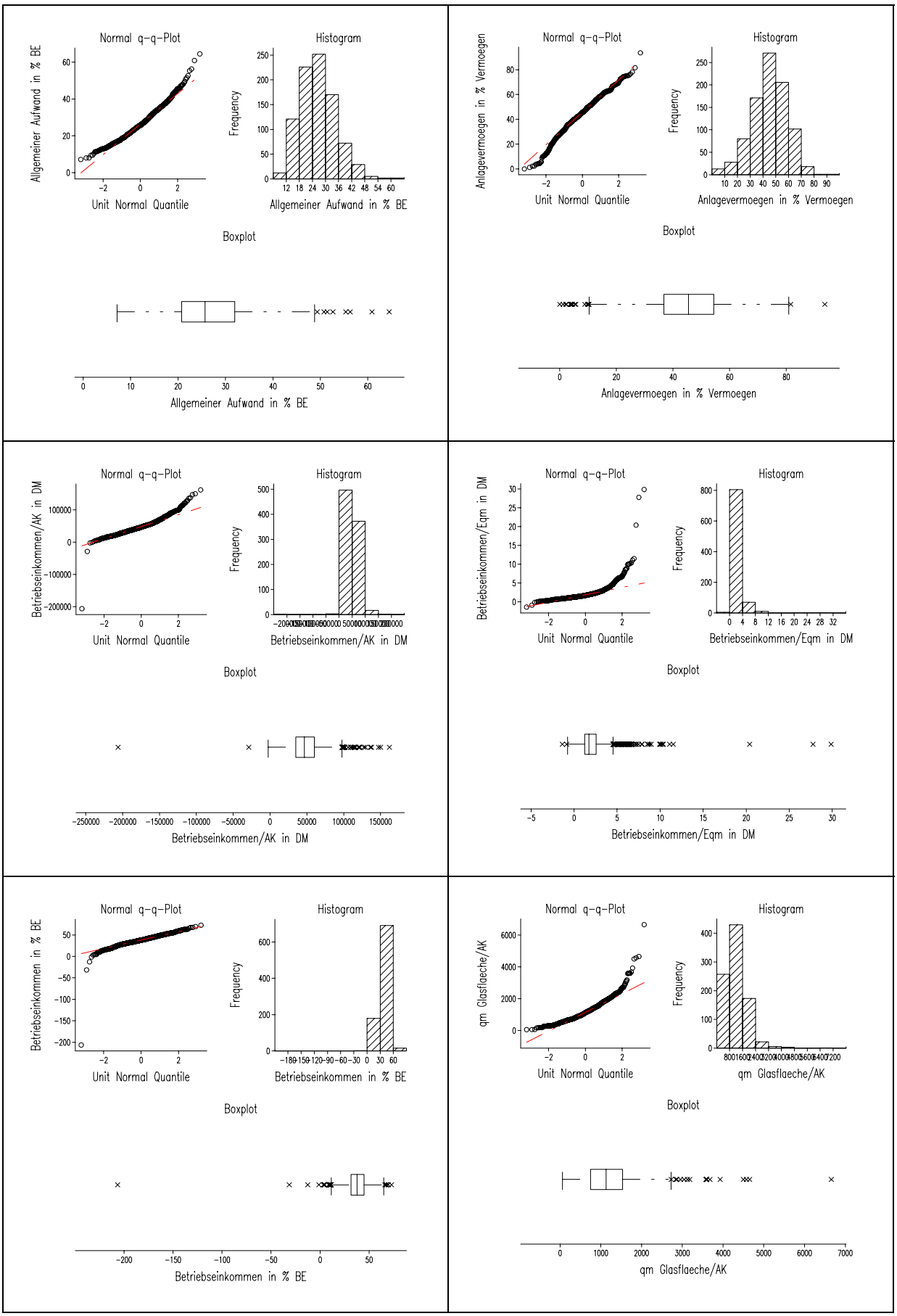


Abbildung B1: Univariate Graphiken zur Beurteilung von Lage- und Dispersionsparametern sowie Verteilungen ausgewählter Kennzahlen

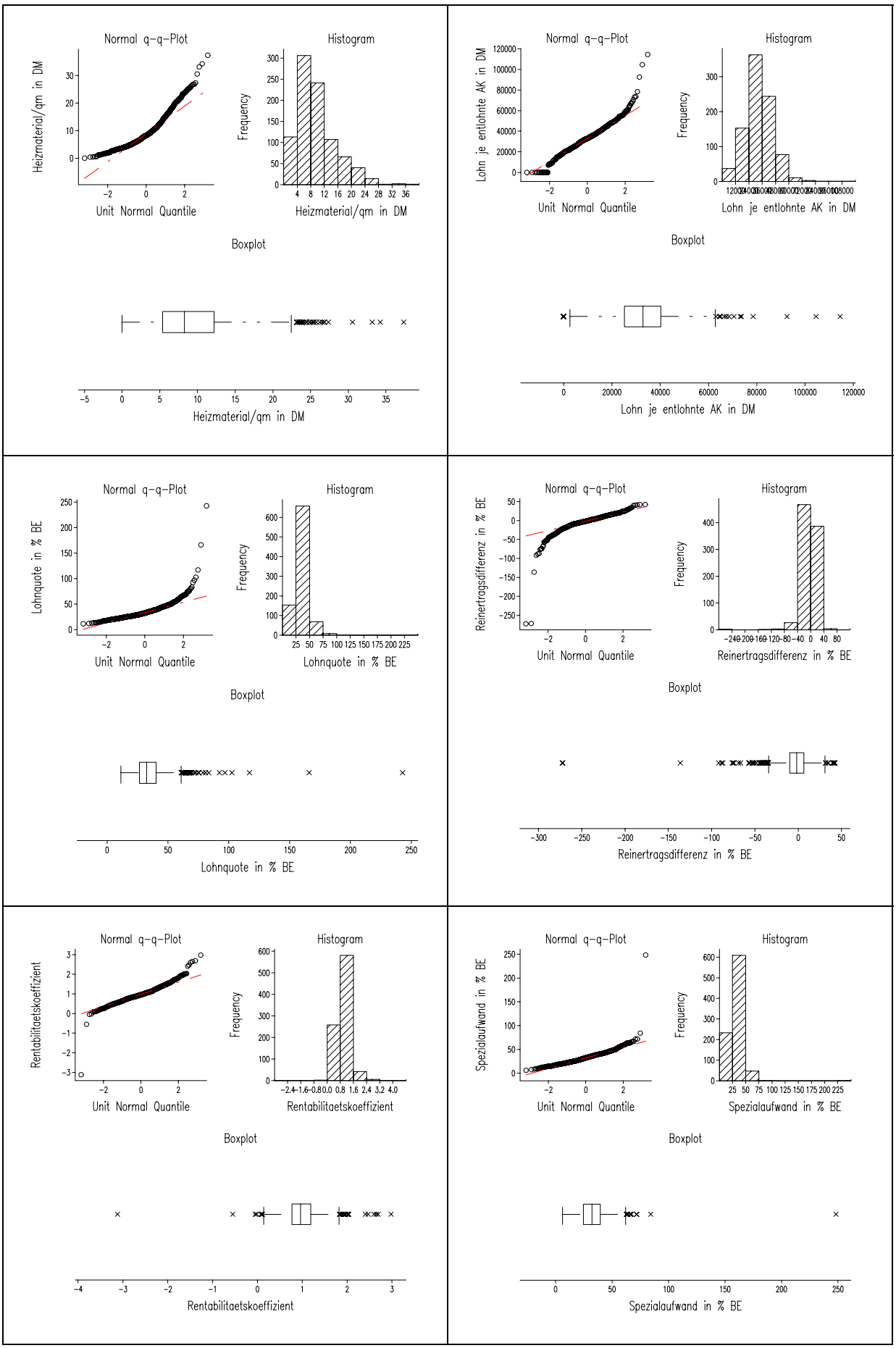


Abbildung B2: Univariate Graphiken zur Beurteilung von Lage- und Dispersionsparametern sowie Verteilungen ausgewählter Kennzahlen

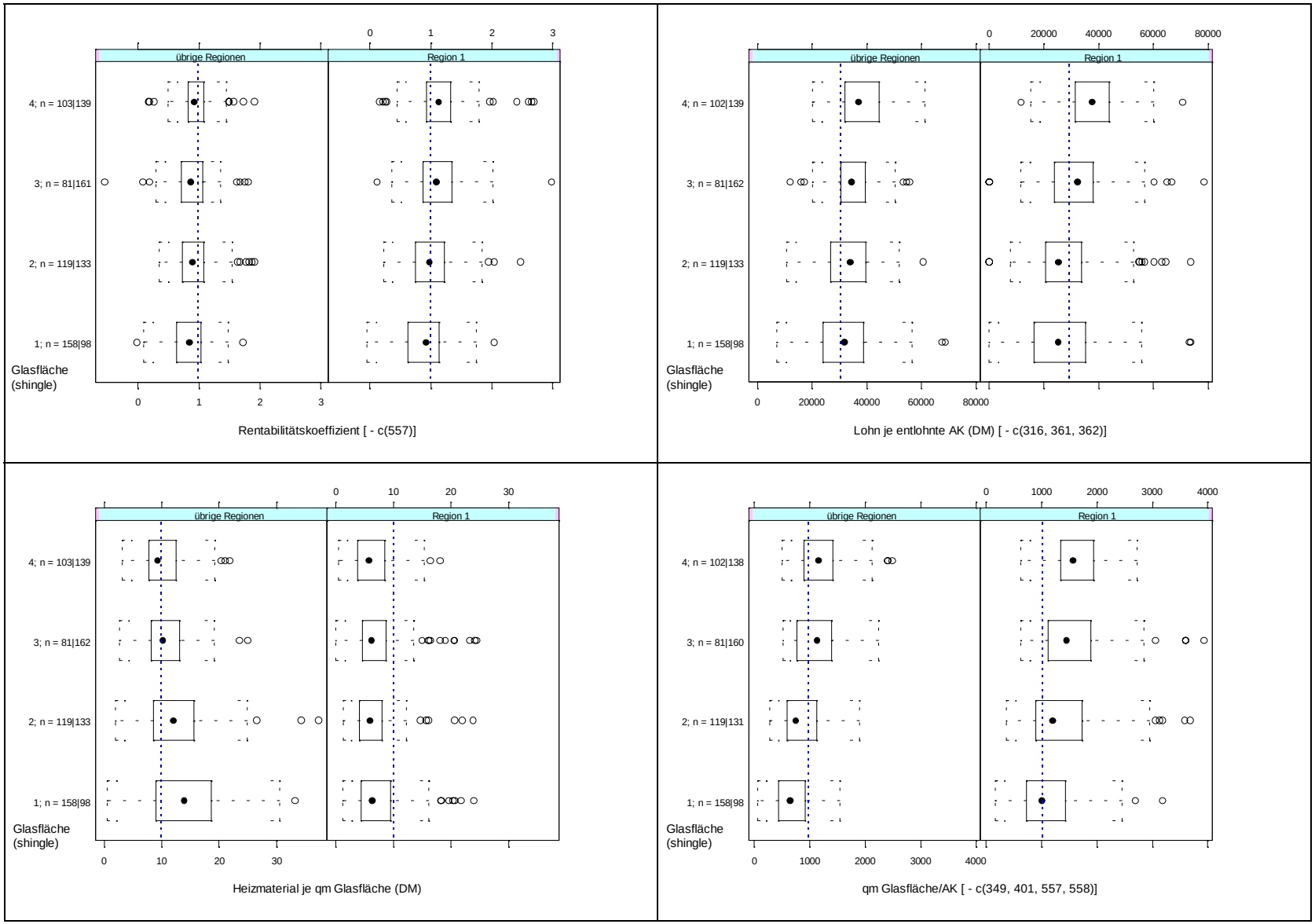
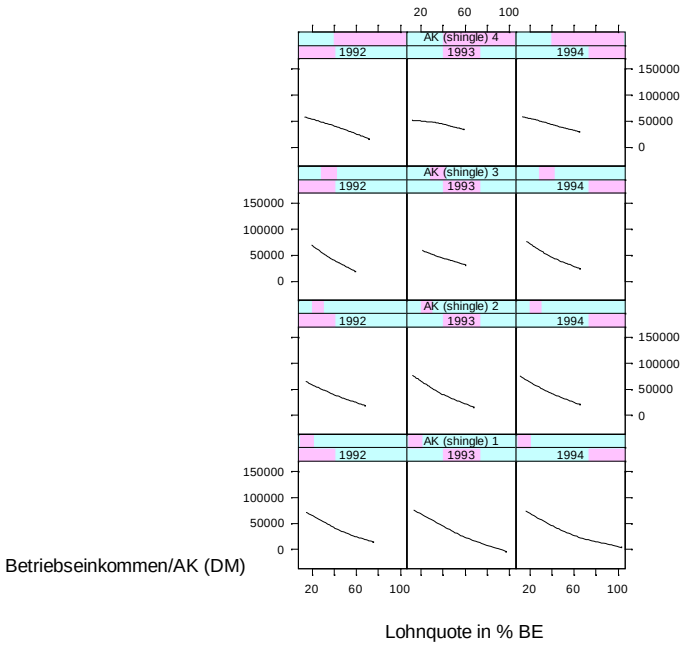
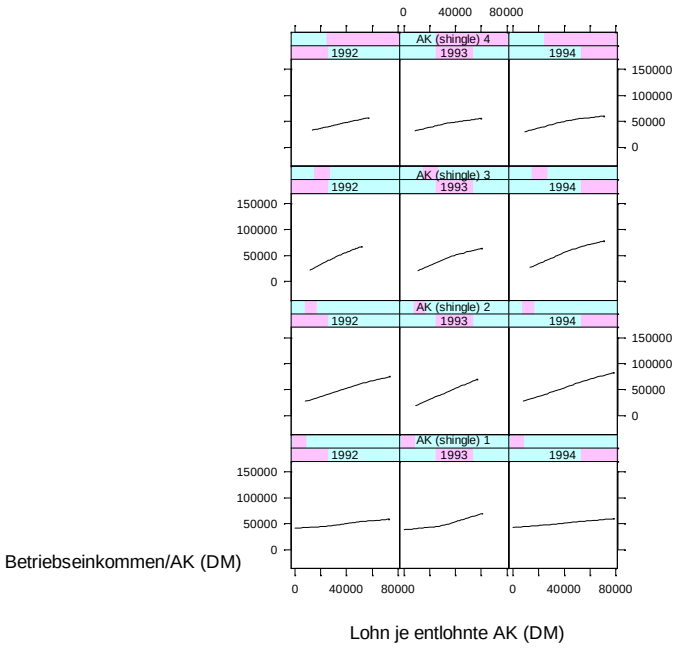


Abbildung B3: Trellis-Displays mit Boxplots für die Kennzahlen Rentabilitätskoeffizient, Lohn je entlohnte AK, Heizmaterial je qm und Glasfläche je AK; konditioniert nach Regionen

a) Betriebseinkommen/AK versus Lohnquote, nach Ausschluß der Betriebe 484, 485, 486, 557



a) Betriebseinkommen/AK versus Lohn je entlohnte AK, nach Ausschluß der Betriebe 316, 361, 362, 557

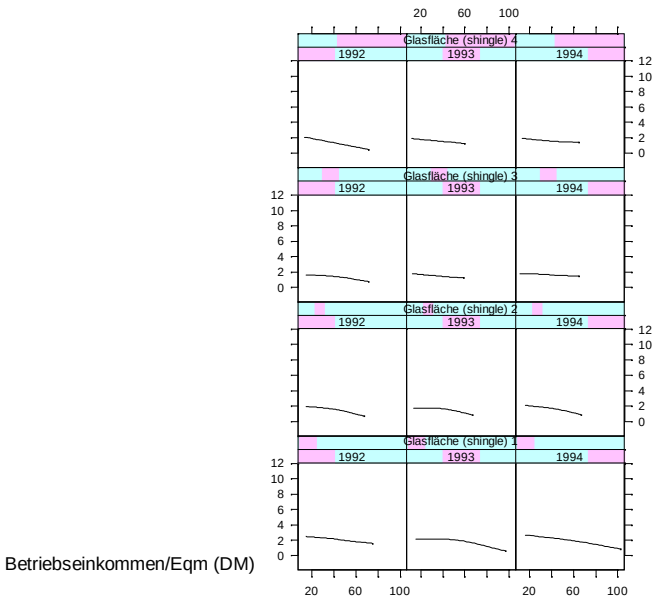


Anzahl Falle in den einzelnen Panels

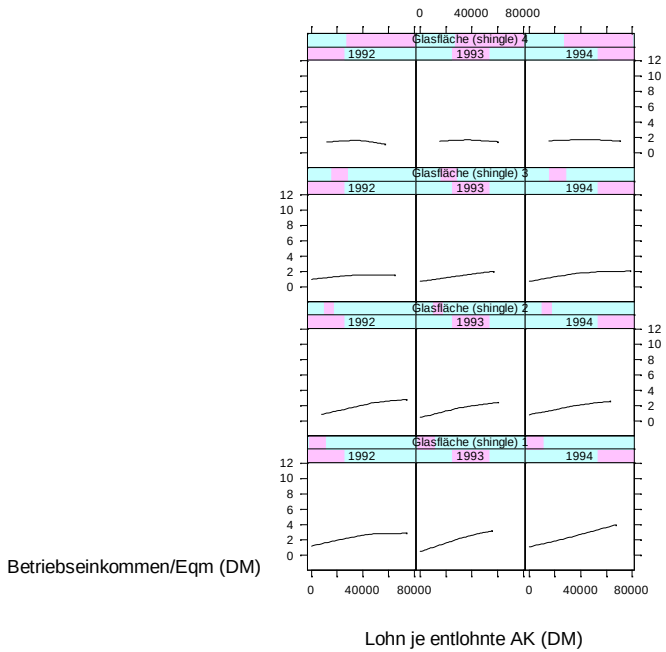
a)			b)		
82	78	81	82	78	81
79	85	79	78	83	78
86	76	82	86	76	82
76	85	79	77	86	81

Abbildung B4: Trellis-Displays mit Loess-Regressionslinien, konditioniert nach Anzahl Arbeitskräfte und Erhebungsjahr, für die Beziehung von Arbeitsproduktivität (Betriebseinkommen/AK) und Lohnquote (a)) beziehungsweise Lohn je entlohnte AK (b))

a) Betriebseinkommen/Eqm versus Lohnquote, nach Ausschluß der Betriebe 484, 485, 486 811, 812, 813



a) Betriebseinkommen/Eqm versus Lohn je entlohnte AK, nach Ausschluß der Betriebe 316, 361, 362, 811, 812, 813

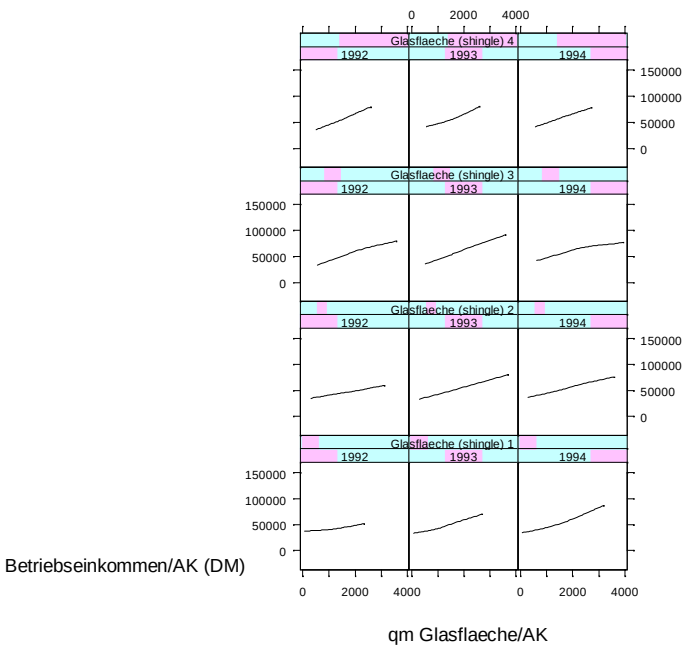


Anzahl Falle in den einzelnen Panels

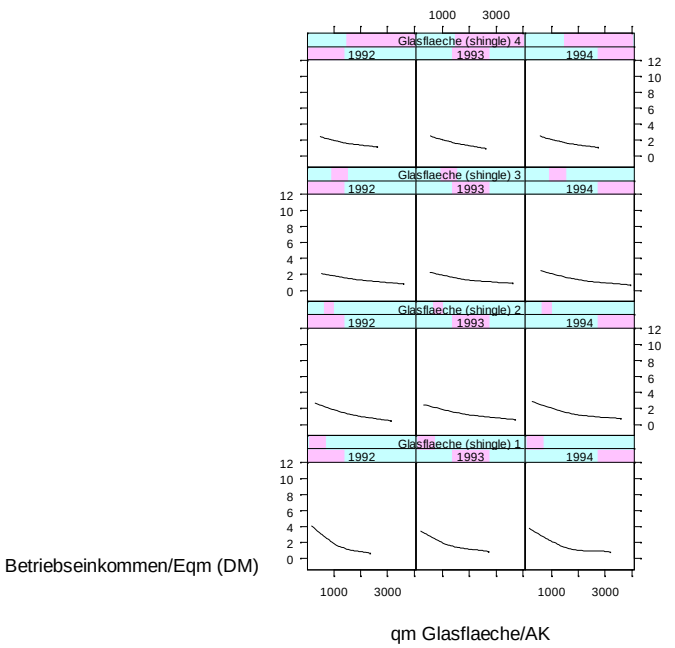
a)			b)		
77	81	84	77	81	83
83	81	79	82	80	79
84	83	85	84	83	85
89	82	79	90	83	81

Abbildung B5: Trellis-Displays mit Loess-Regressionslinien, konditioniert nach Shingle Glasfläche und Erhebungsjahr, für die Beziehung von Flächenproduktivität (Betriebseinkommen/Eqm) und Lohnquote (a)) beziehungsweise Lohn je entlohnte AK (b))

a) Betriebseinkommen/AK versus Glasfläche/AK, nach Ausschluß der Betriebe 349, 401, 557, 558



b) Betriebseinkommen/Eqm versus Glasfläche/AK nach Ausschluß der Betriebe 349, 401, 557, 558, 811, 812, 813

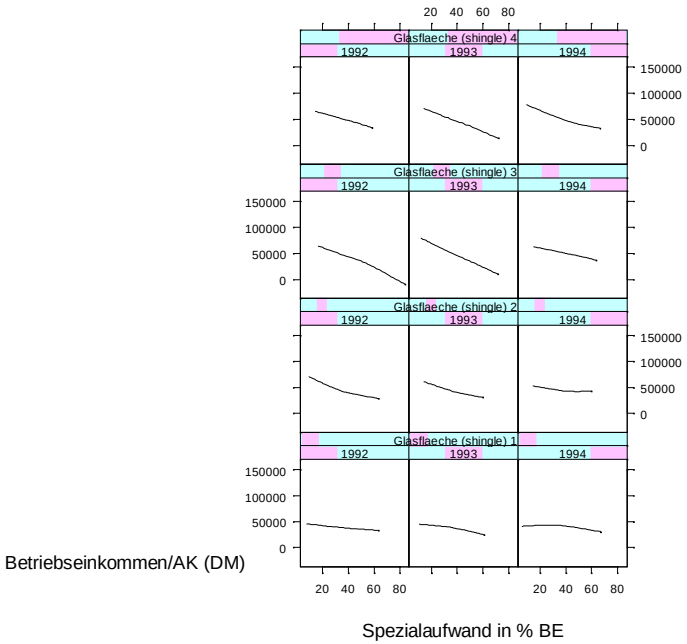


Anzahl Falle in den einzelnen Panels

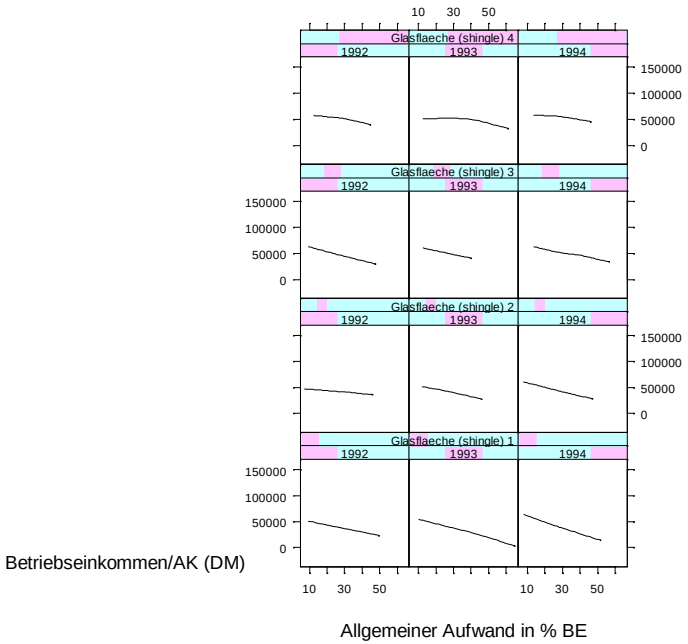
a)			b)		
77	81	84	77	81	84
82	80	79	82	80	79
83	82	85	83	82	85
91	84	81	90	83	80

Abbildung B6: Trellis-Displays mit Loess-Regressionslinien, konditioniert nach Shingle Glasfläche und Erhebungsjahr, für die Beziehung von Arbeitsproduktivität (Betriebseinkommen/AK) (a)) und Flächenproduktivität (Betriebseinkomen/Eqm) (b)) zu qm Glasfläche/AK

a) Betriebseinkommen/AK versus Spezialaufwand, nach Ausschluß des Betriebes 557



b) Betriebseinkommen/AK versus Allgemeiner Aufwand nach Ausschluß des Betriebes 557

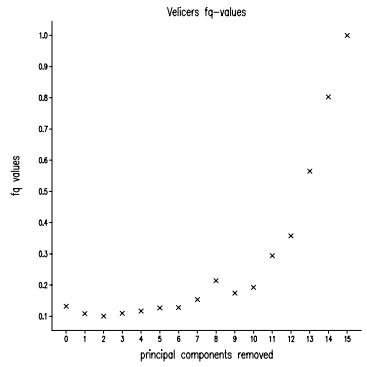


Anzahl Falle in den einzelnen Panels

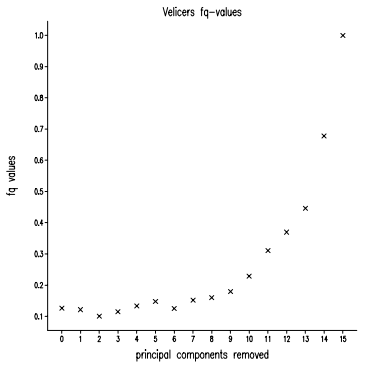
a)			b)		
77	81	84	77	81	84
83	80	79	83	80	79
84	83	85	84	83	85
91	84	81	91	84	81

Abbildung B7: Trellis-Displays mit Loess-Regressionslinien, konditioniert nach Shingle Glasfläche und Erhebungsjahr, für die Beziehung von Arbeitsproduktivität (Betriebseinkommen/AK) zu Spezialaufwand (a)) und allgemeinem Aufwand (b))

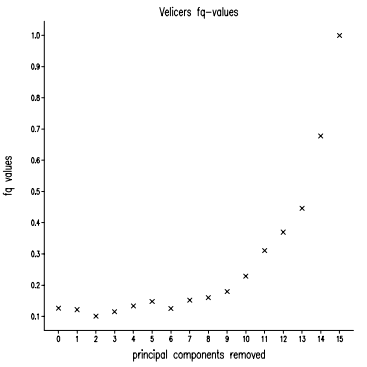
1



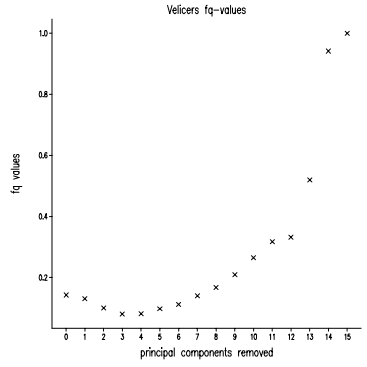
2



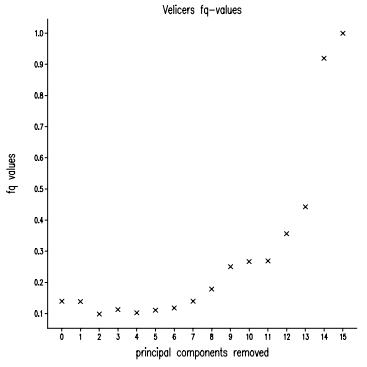
3



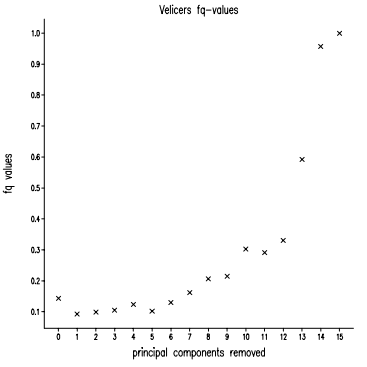
4



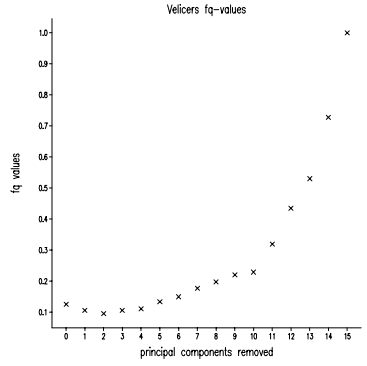
5



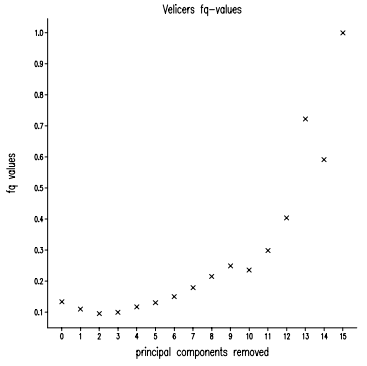
6



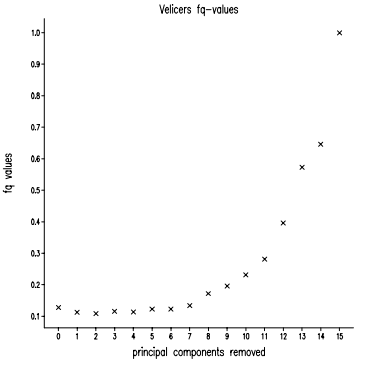
7



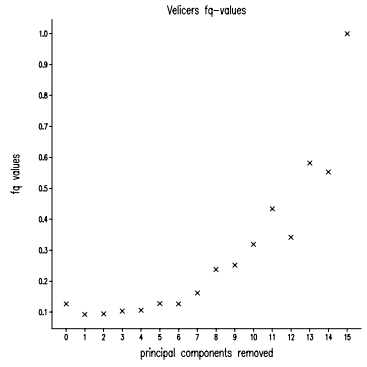
8



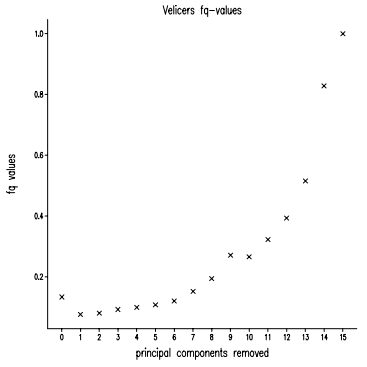
9



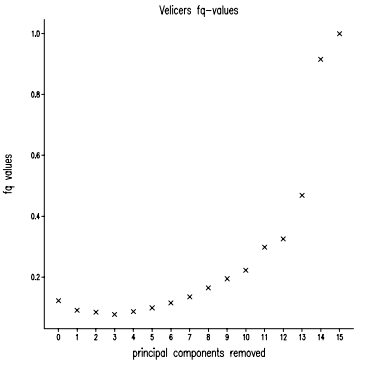
10



11



12



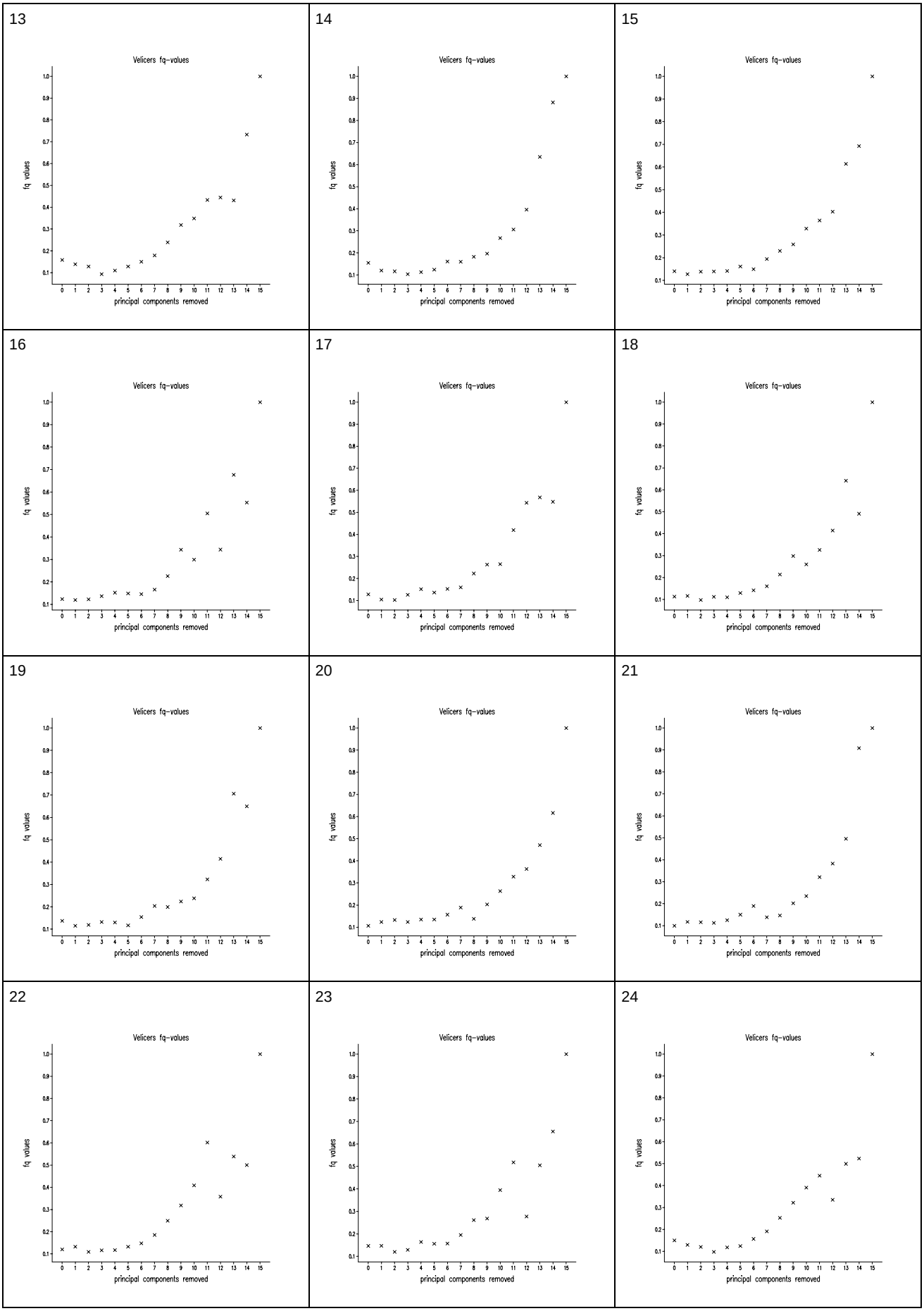
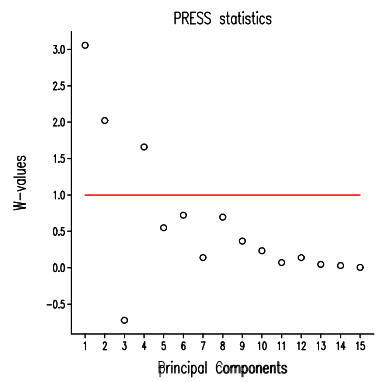
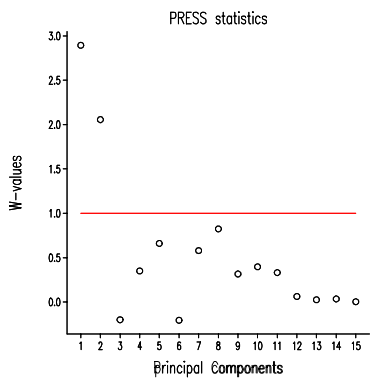


Abbildung B8: f_q-Werte zur Bestimmung der Anzahl der ‘wesentlichen’ Hauptkomponenten in den 24 Gruppen der Kennzahlenbetriebe

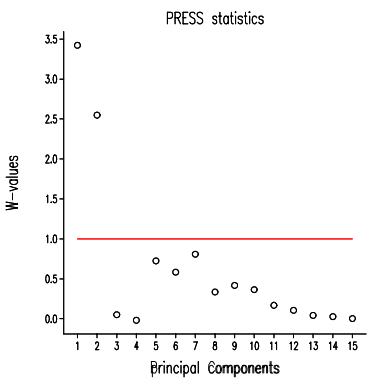
1



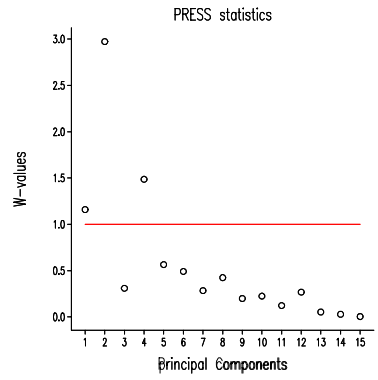
2



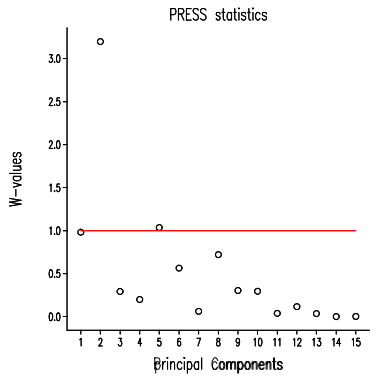
3



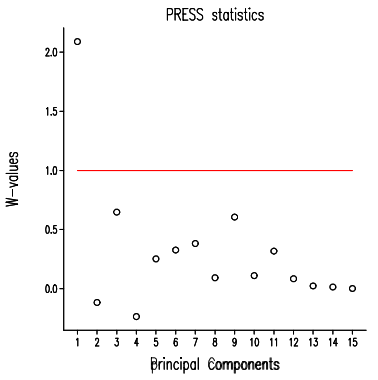
4



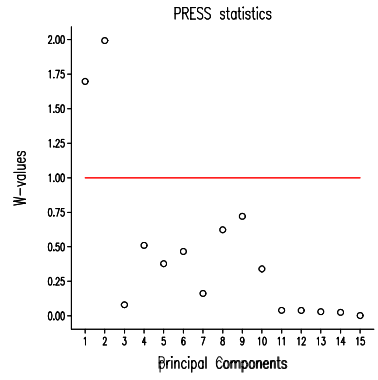
5



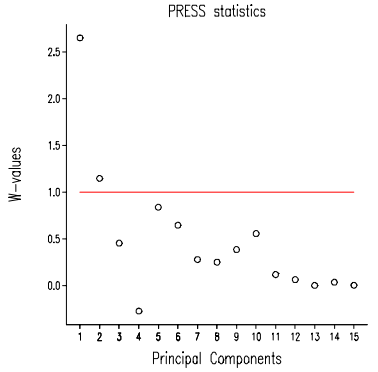
6



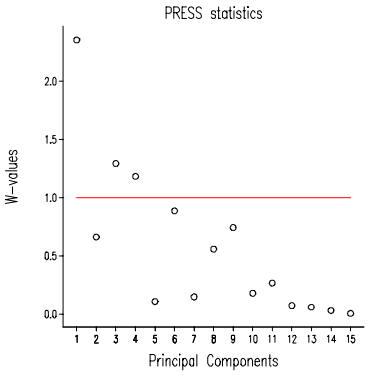
7



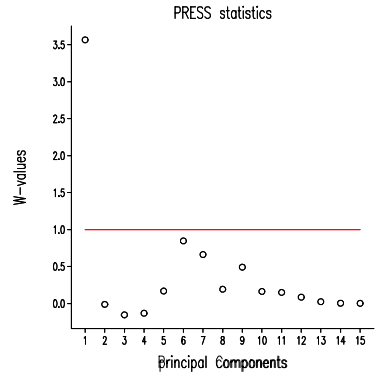
8



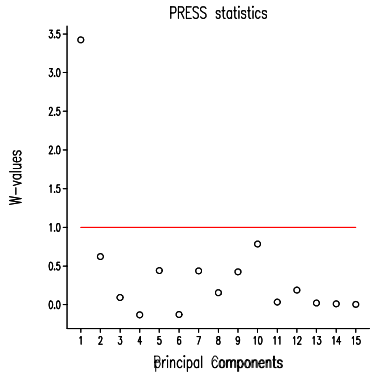
9



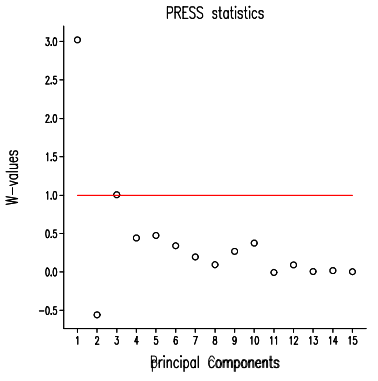
10



11



12



13

14

15

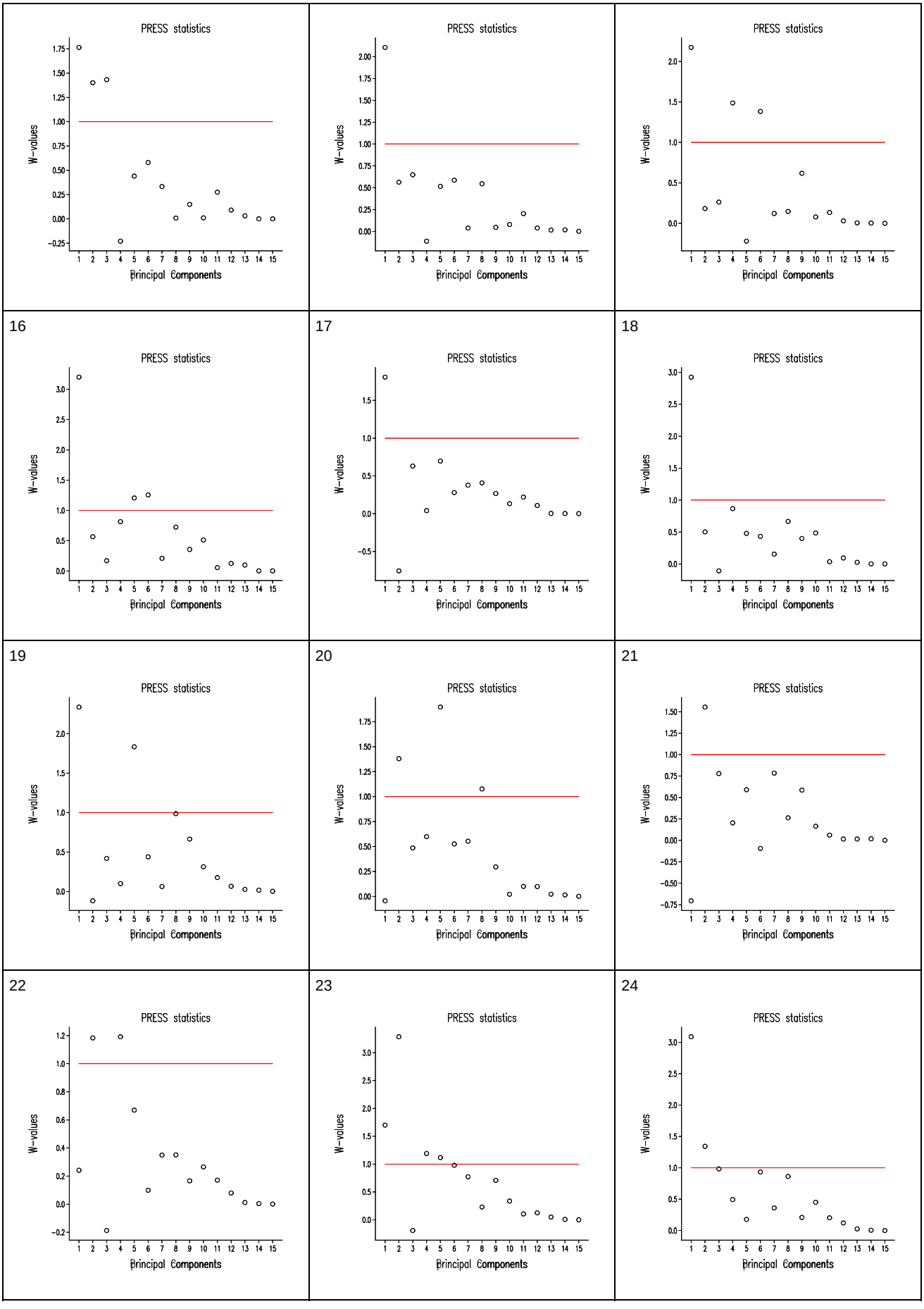


Abbildung B9: W-Werte zur Bestimmung der Anzahl der ‘wesentlichen’ Hauptkomponenten in den 24 Gruppen der Kennzahlenbetriebe

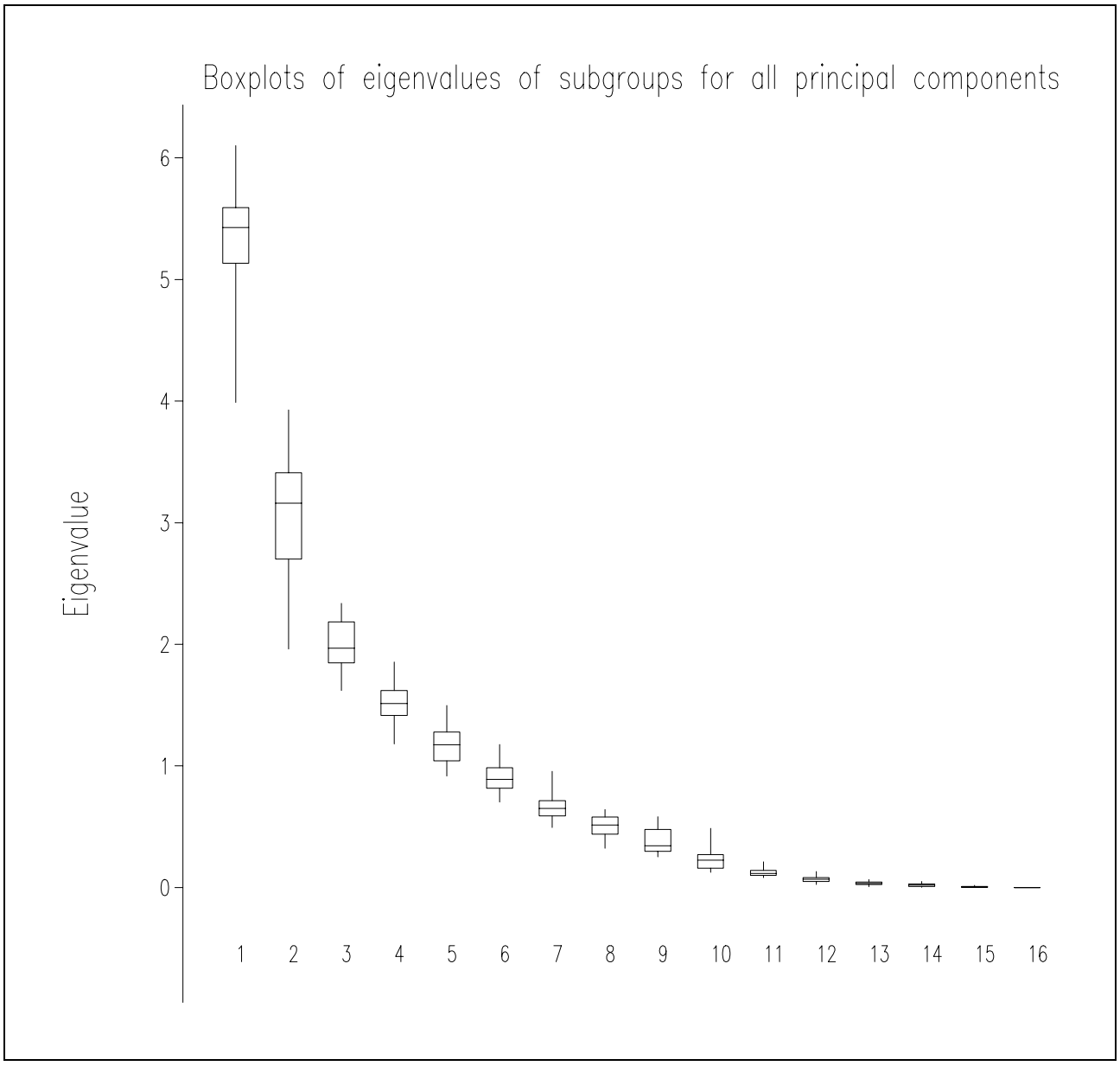


Abbildung B10: Boxplots der Eigenwerte der Hauptkomponentenanalysen aller 24 Gruppen der Kennzahlenbetriebe

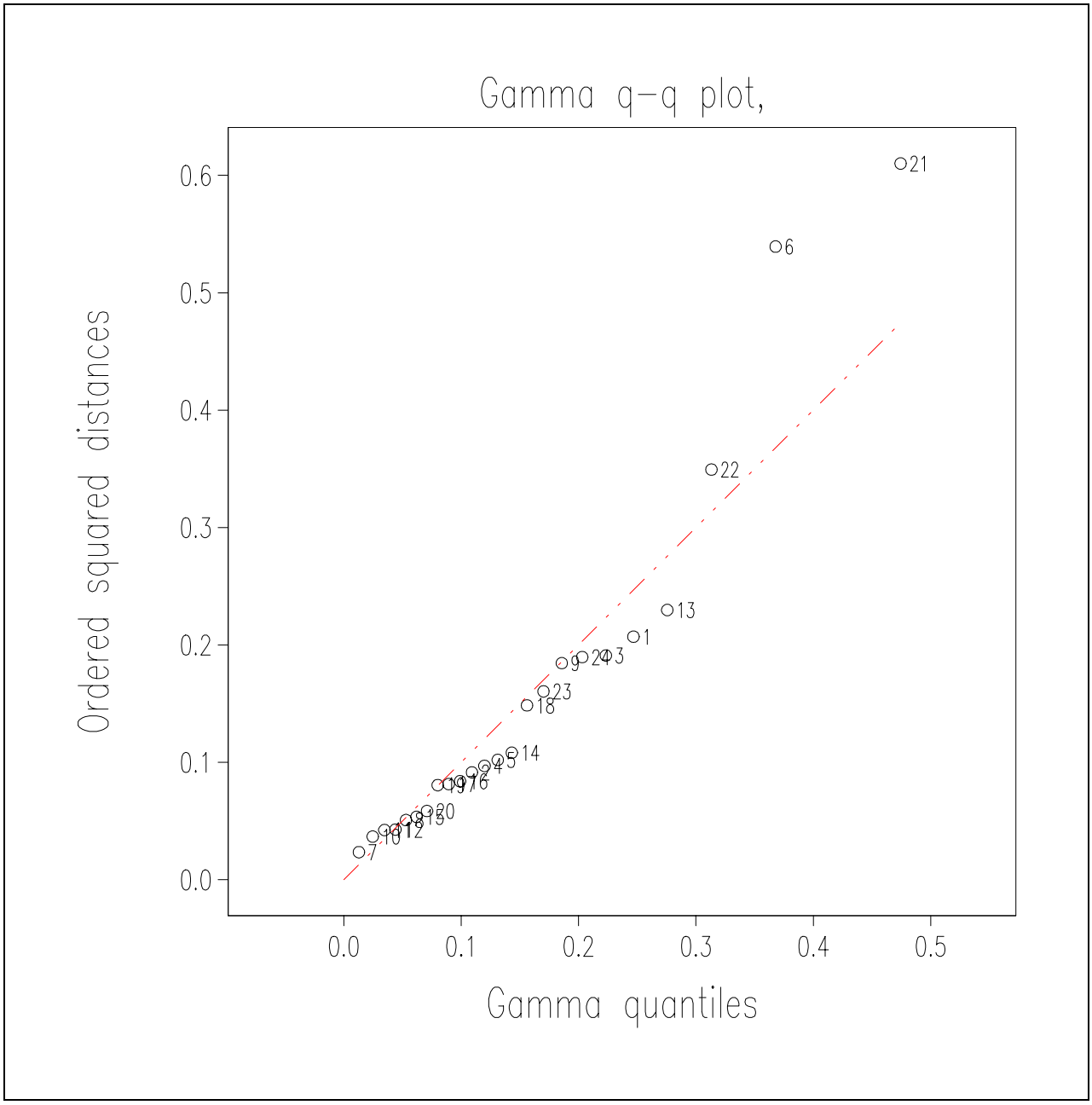


Abbildung B11: Gamma q-q-Plot für den Vergleich des ersten Eigenvektors der 24 Gruppen der Kennzahlenbetriebe mit dem ‘typischen’ ersten Eigenvektor

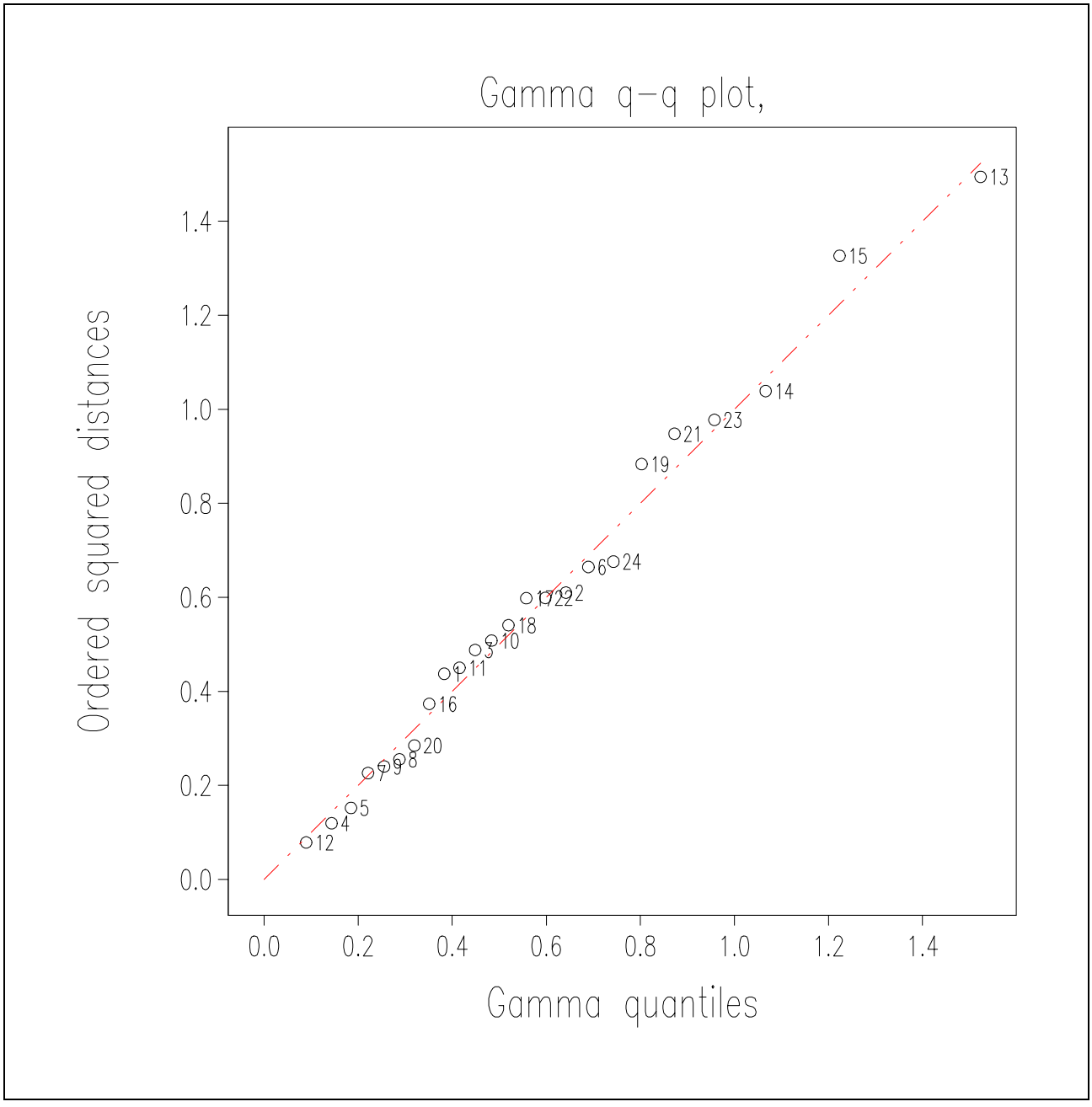


Abbildung B12: Gamma q-q-Plot für den Vergleich des zweiten Eigenvektors der 24 Gruppen der Kennzahlenbetriebe mit dem ‘typischen’ ersten Eigenvektor

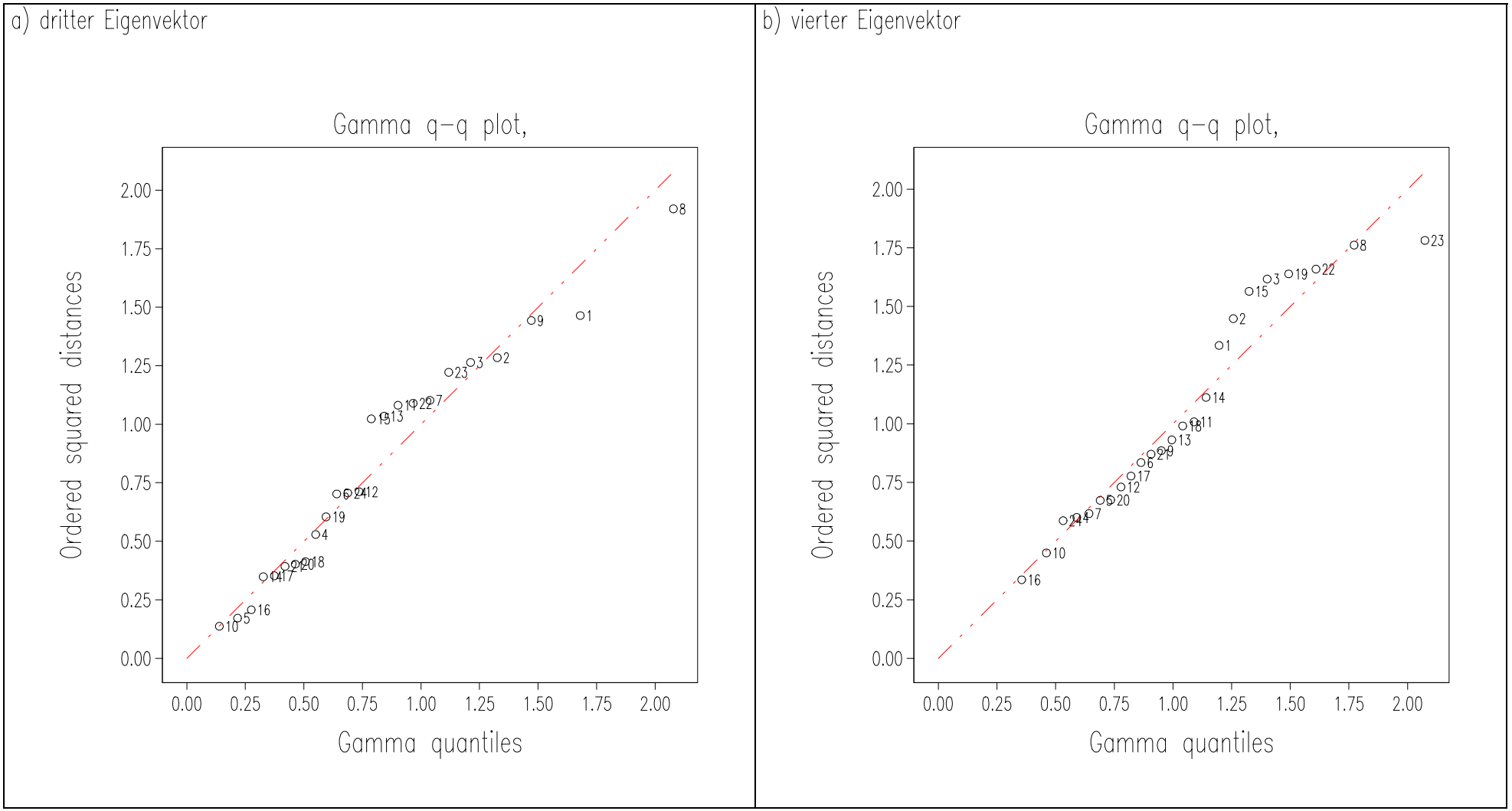


Abbildung B13: Gamma q-q-Plot für den Vergleich des dritten (a)) und vierten (b)) Eigenvektors der 24 Gruppen der Kennzahlenbetriebe mit dem ‘typischen’ ersten Eigenvektor

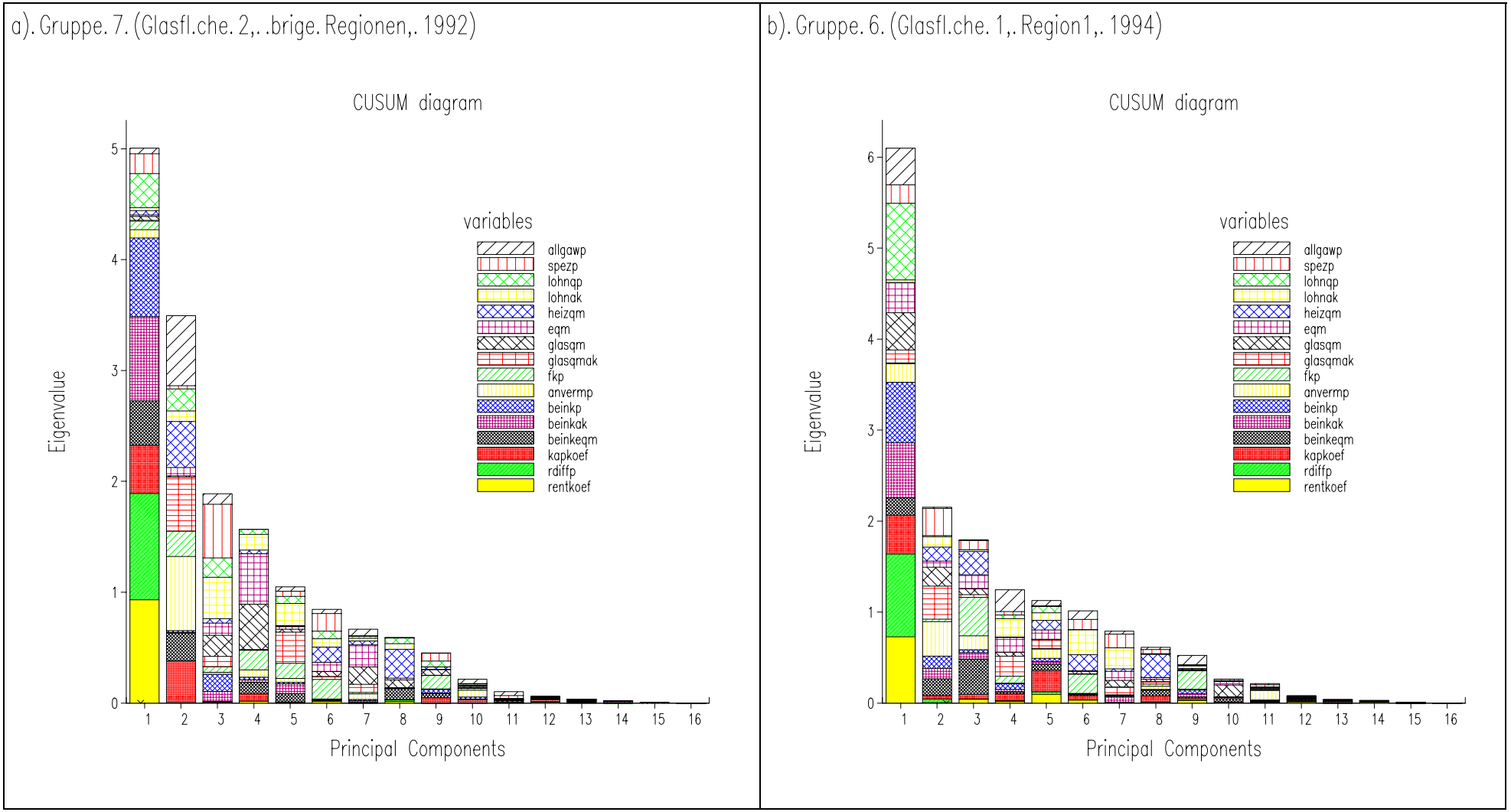


Abbildung B14a: CUSUM-Diagramme für Gruppe 7 (a)), Gruppe 6 (b)) der 24 Gruppen

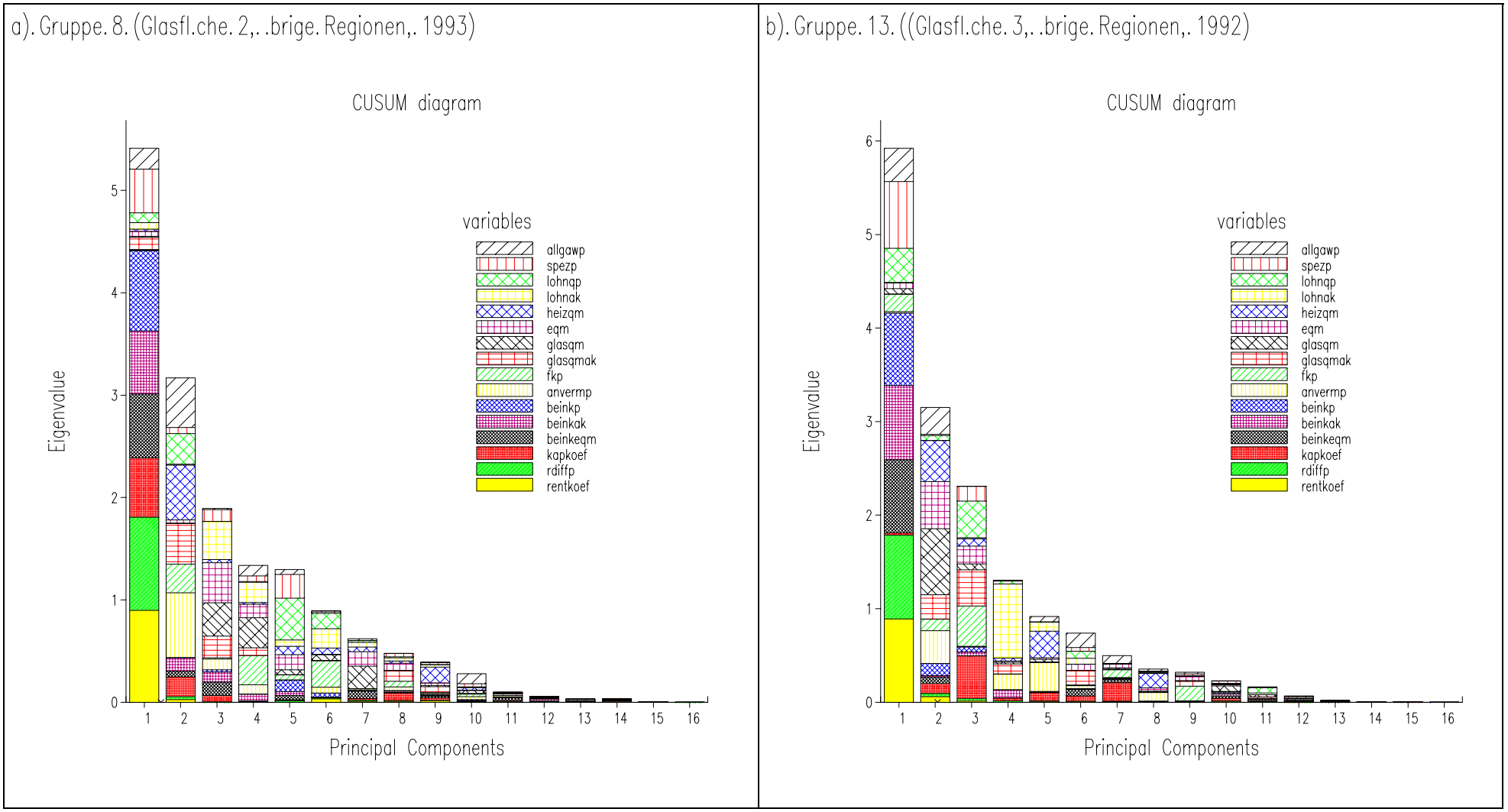


Abbildung 14b: CUSUM-Diagramme für Gruppe 8)) und Gruppe 13)) der 24 Gruppen

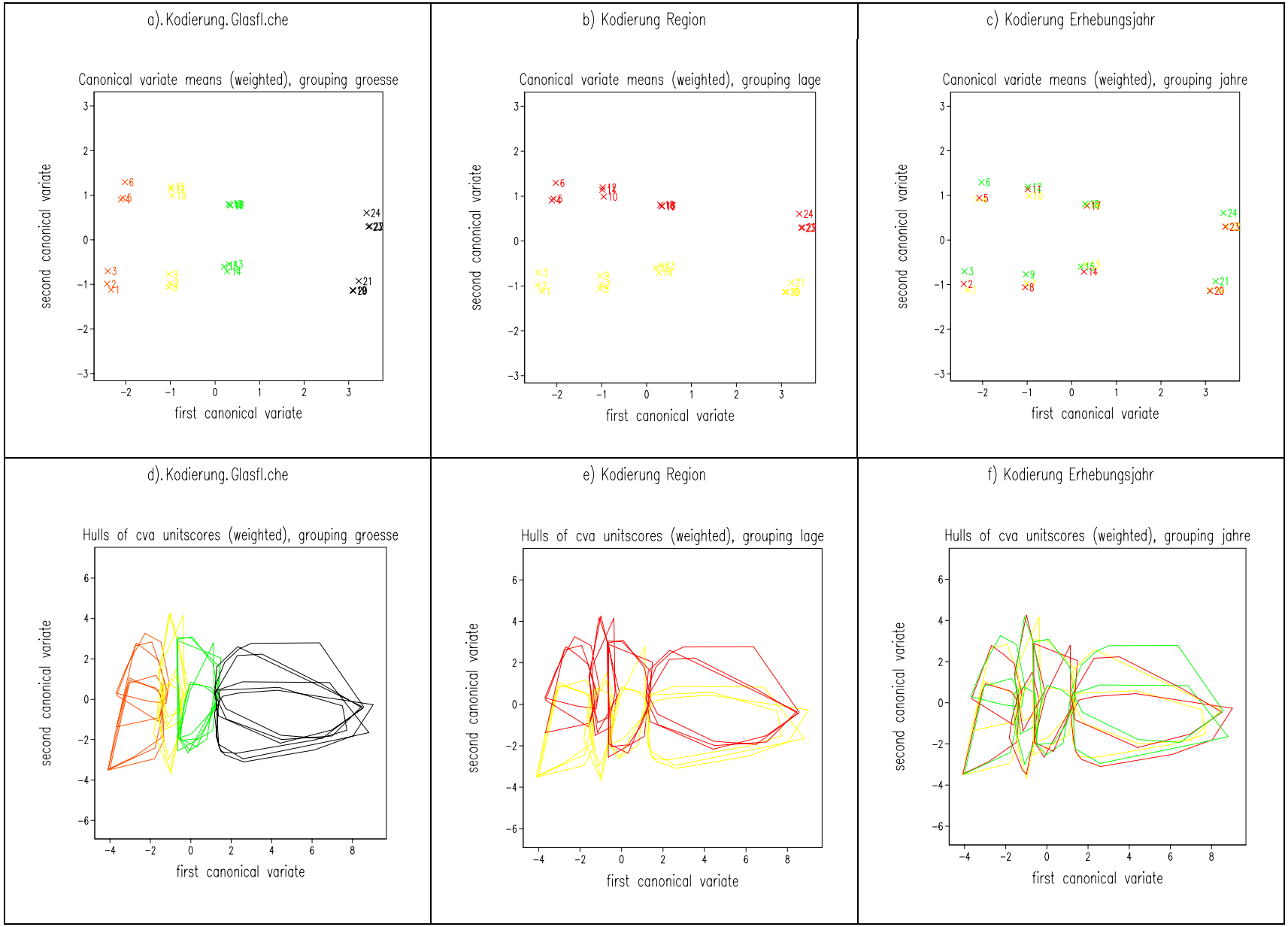


Abbildung B15: Gewichtete CVA-Mittelwerte und konvexe Hüllen der Objektkonfigurationen, farblich kodiert nach Erhebungsjahr, Glasfläche und Region; Anteil erklärter Varianz durch die erste Dimension 77,6%, durch die zweite Dimension 14,3%

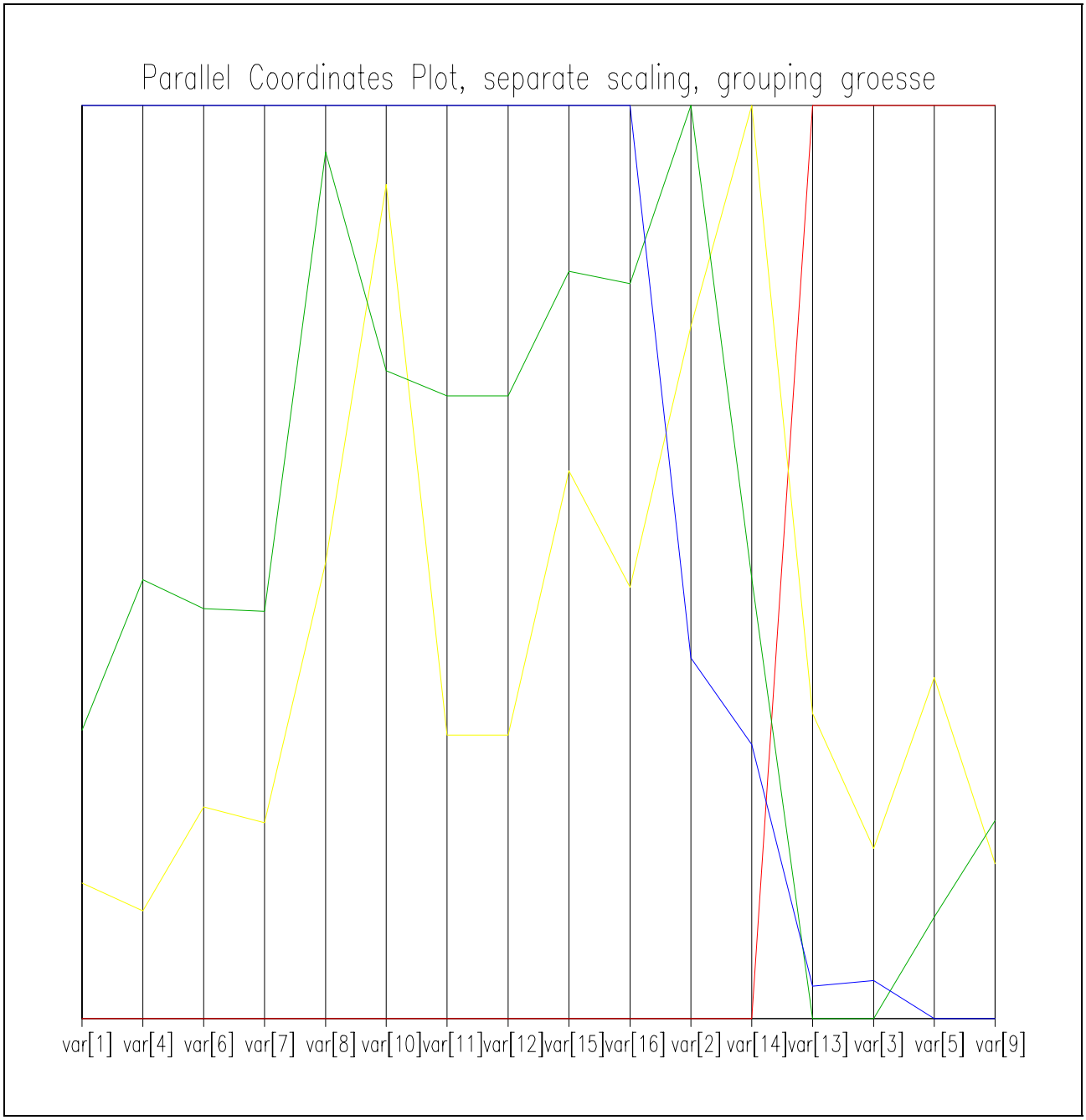


Abbildung B16: Parallekoordinatenplot der Orginalwerte der in der kanonischen Variablenanalyse verrechneten Kennzahlen

var[1]	allgawp	var[11]	beinkp	var[14]	kapkoef
var[4]	lohnak	var[12]	beinkak	var[13]	beinkeqm
var[6]	eqm	var[15]	rdiffp	var[3]	lohnqp
var[7]	glasqm	var[16]	rentkoef	var[5]	heizqm
var[8]	glasqmak	var[2]	spezp	var[9]	fkp
var[10]	anvermp				

Glasfläche bis 3300 m²
Glasfläche über 3300 m²
bis einschließlich 4980 m²
Glasfläche über 4980 m²
bis einschließlich 7580 m²
Glasfläche über 7580 m²

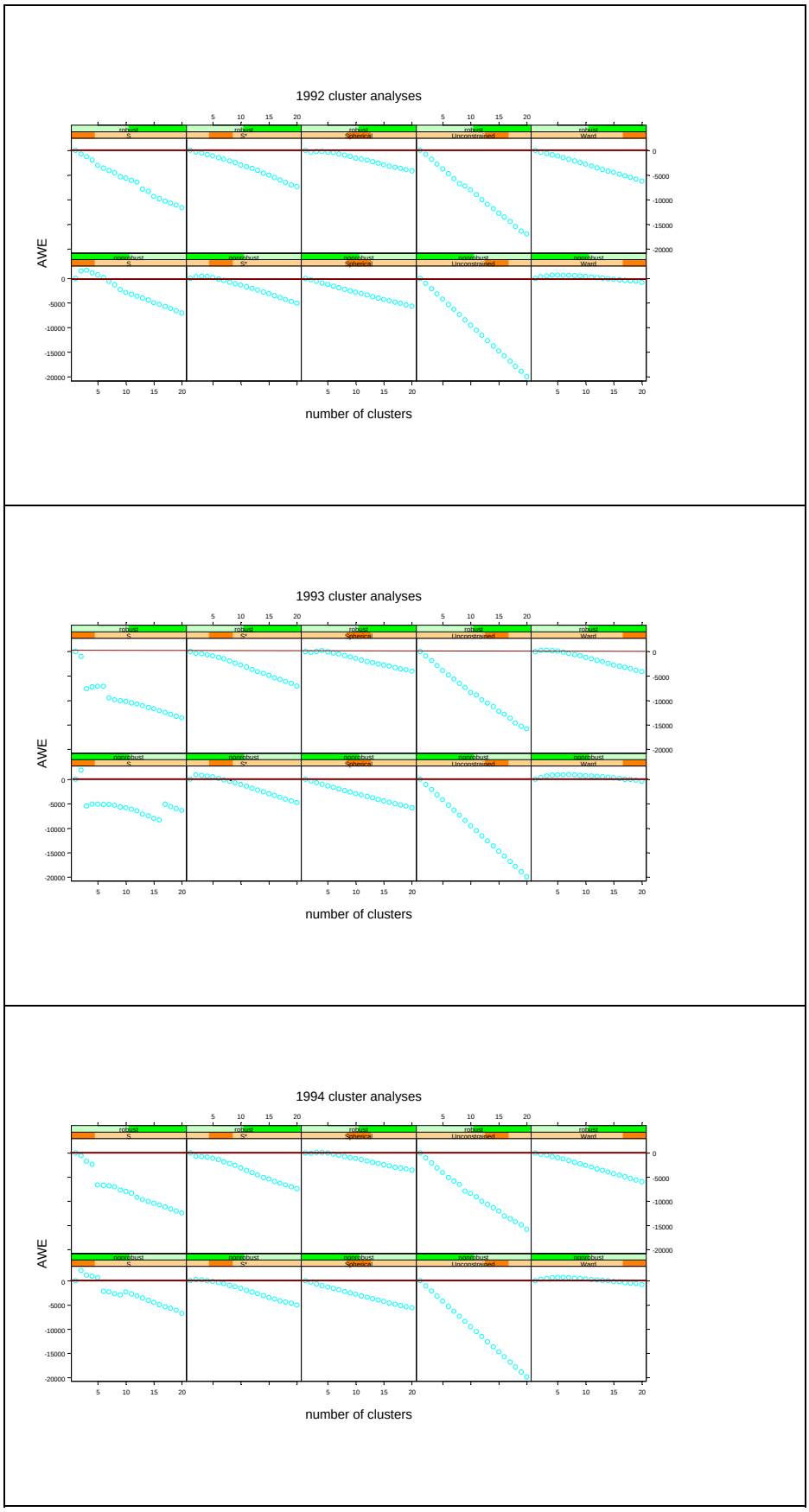


Abbildung B17: AWE-Werte nach verschiedenen Verfahren modellbegründeter Clusteranalyse für 1 bis 20 Cluster und normales und robustes Vorgehen für 1992, 1993 und 1994

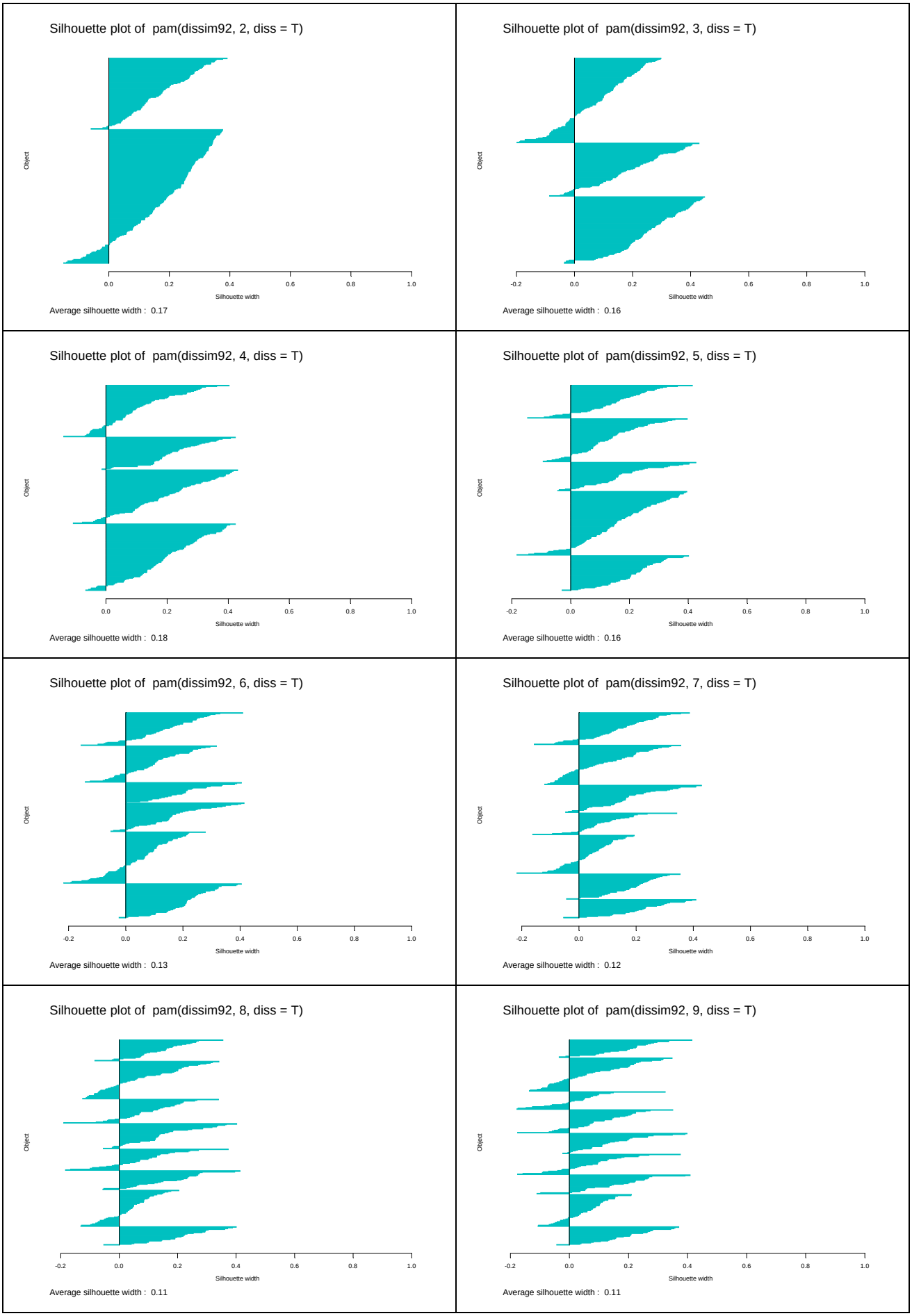


Abbildung B18a: Silhouettenplots für 2 bis 9 Clusterlösungen bei nicht-hierarchischer Klassifikation (Partition um Medoide), 1992

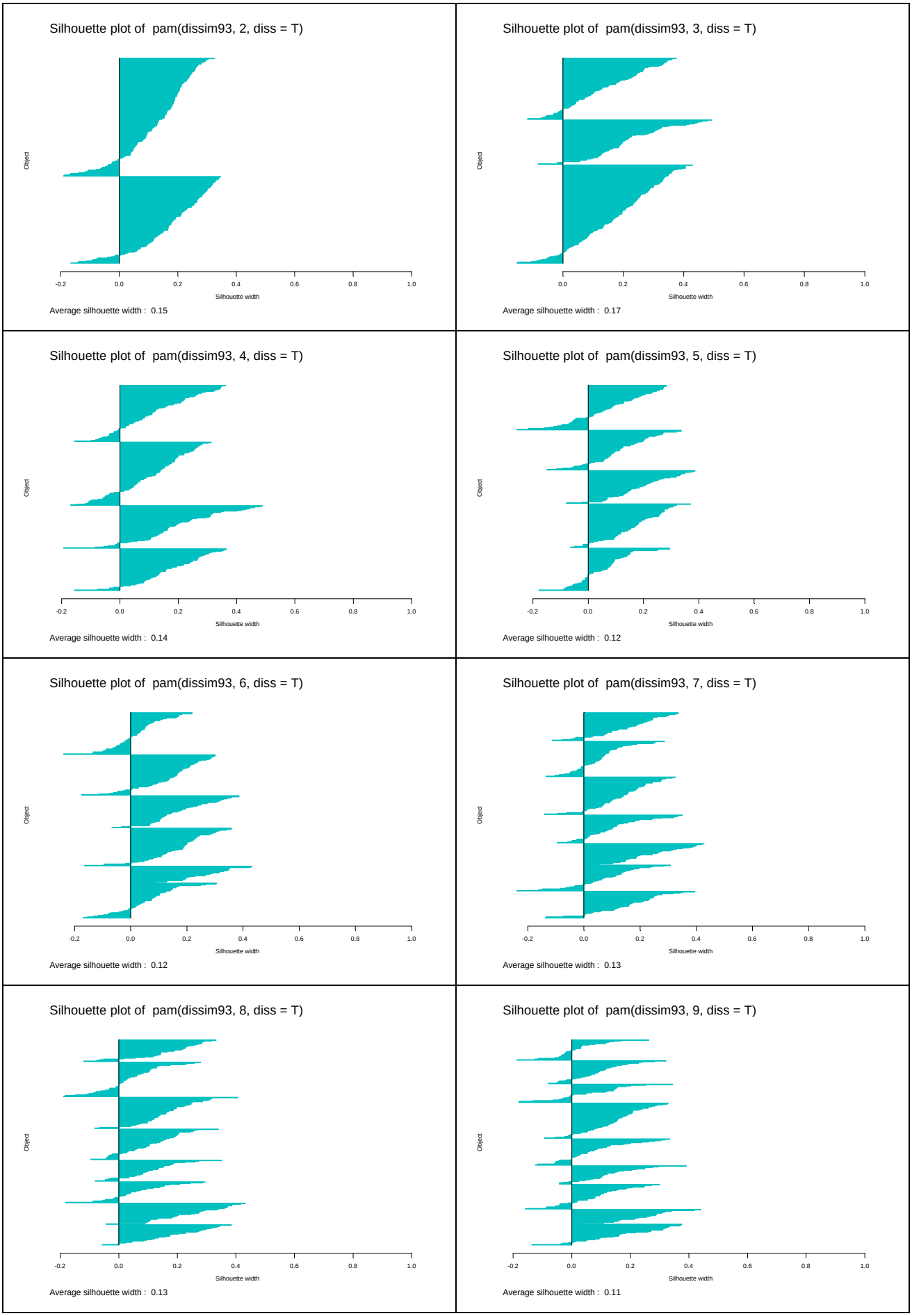


Abbildung B18b: Silhouettenplots für 2 bis 9 Clusterlösungen bei nicht-hierarchischer Klassifikation (Partition um Medoide), 1993

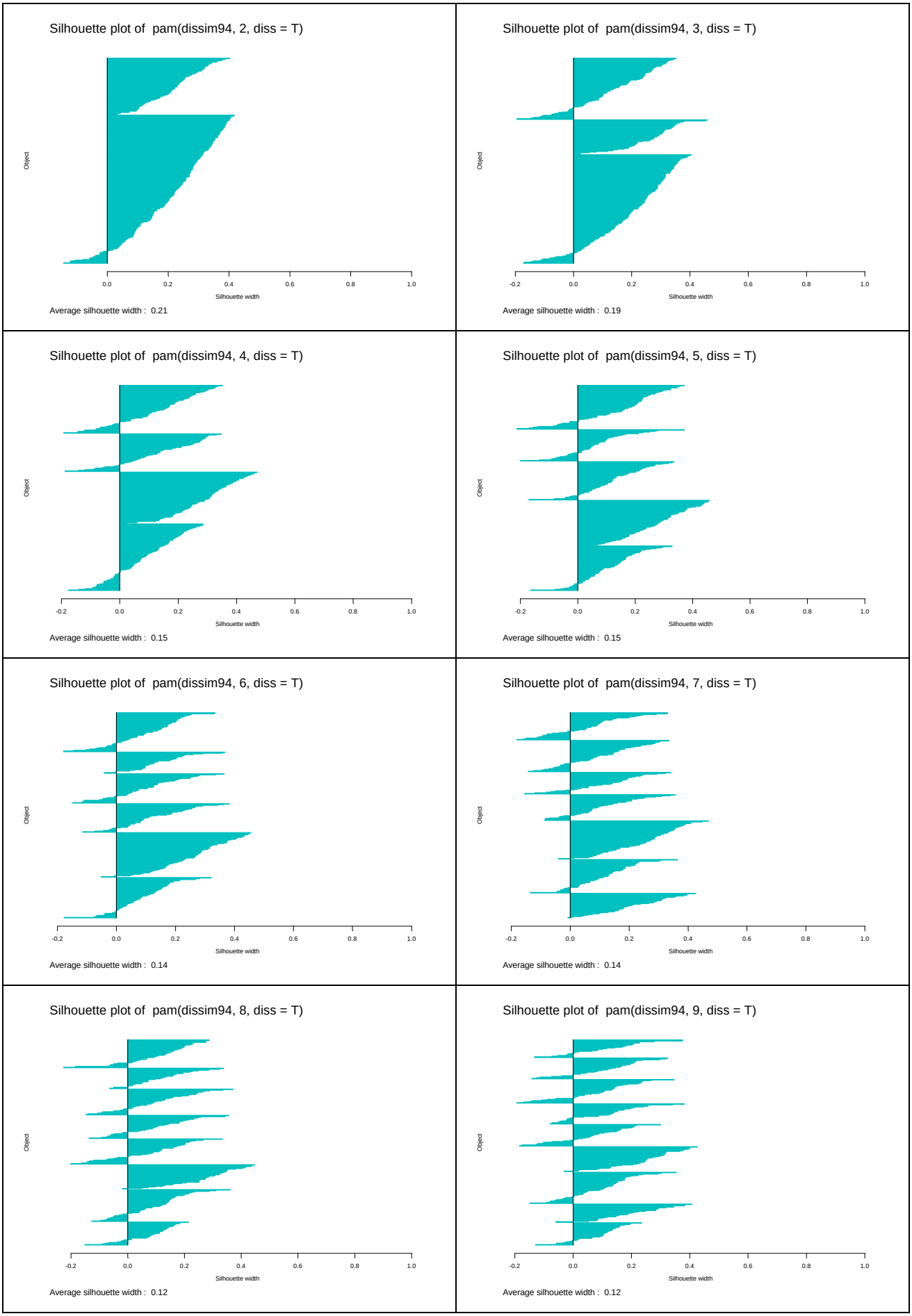
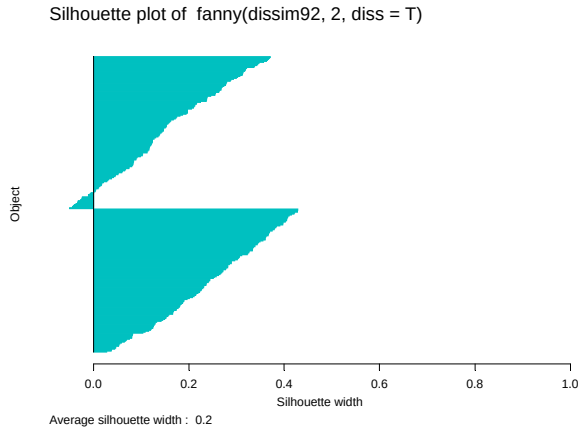
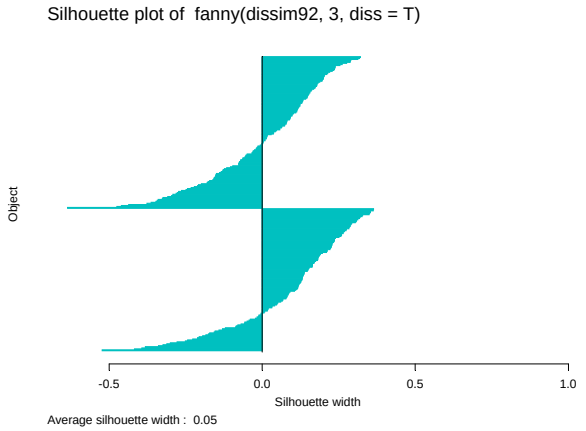
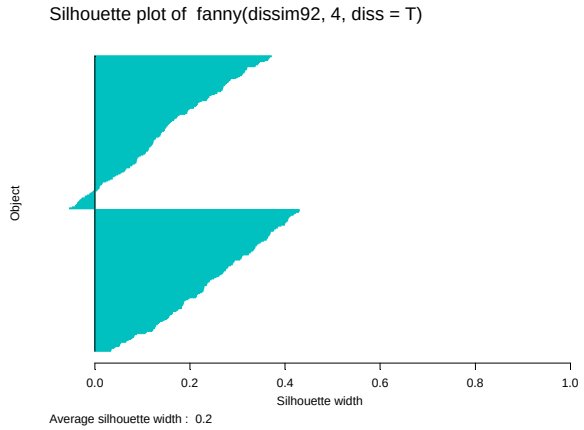
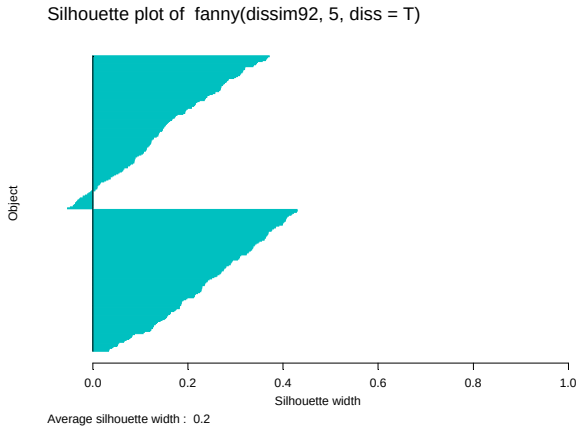


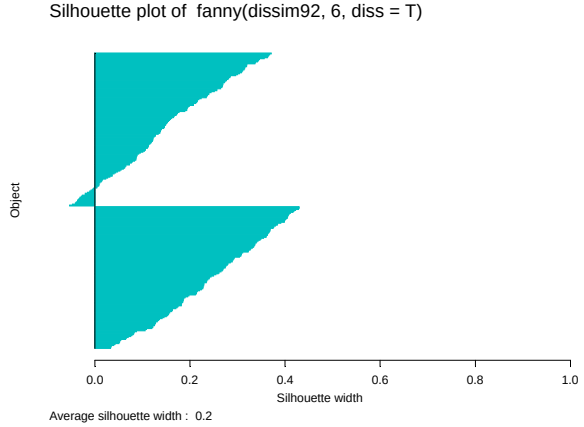
Abbildung B18c: Silhouettenplots für 2 bis 9 Clusterlösungen bei nicht-hierarchischer Klassifikation (Partition um Medoide), 1994





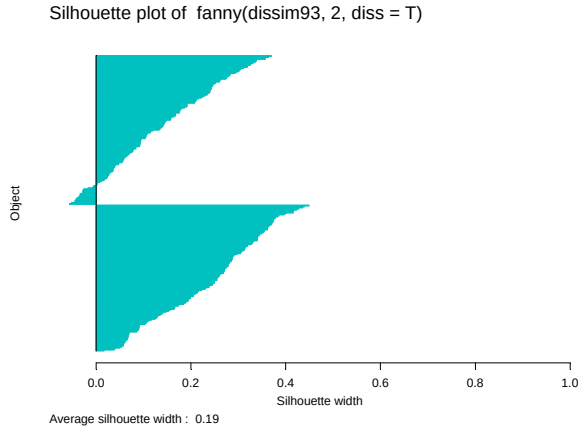


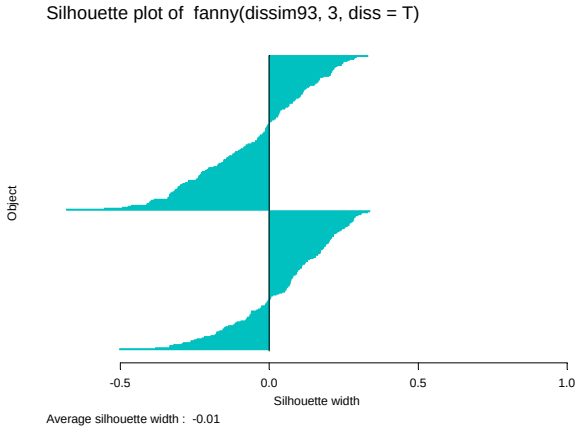


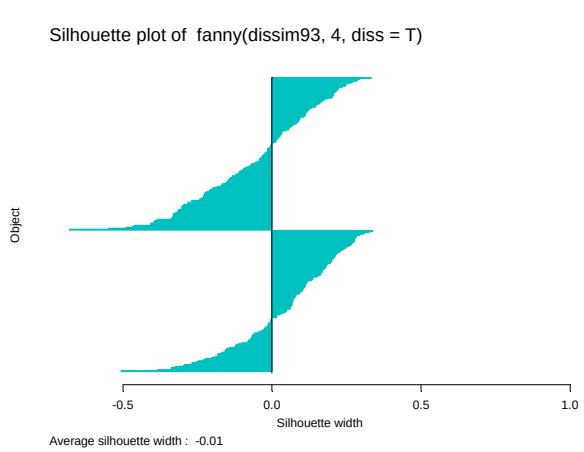


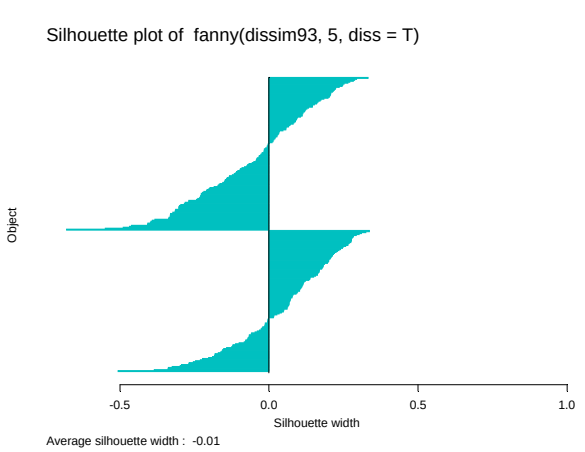

```
2 Cluster
dunn_coeff    normalized
0.5           2.664535e-015
3 Cluster
dunn_coeff    normalized
0.3333333     1.998401e-015
4 Clsuster
dunn_coeff    normalized
0.25          4.440892e-016
5 Cluster
dunn_coeff    normalized
0.2           1.94289e-015
6 Cluster
dunn_coeff    normalized
0.1666667     7.105427e-016
```

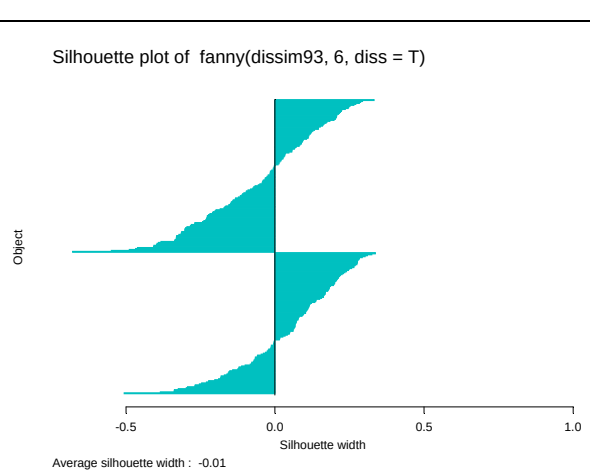
Abbildung B 19a: Silhouettenplots für 2 bis 6 Clusterlösungen bei Fuzzy Clusterung, 1992





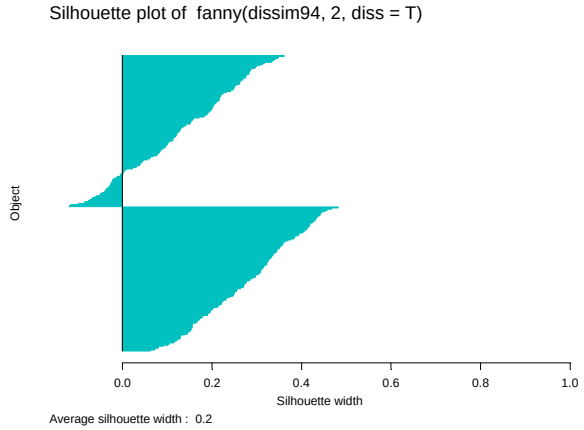


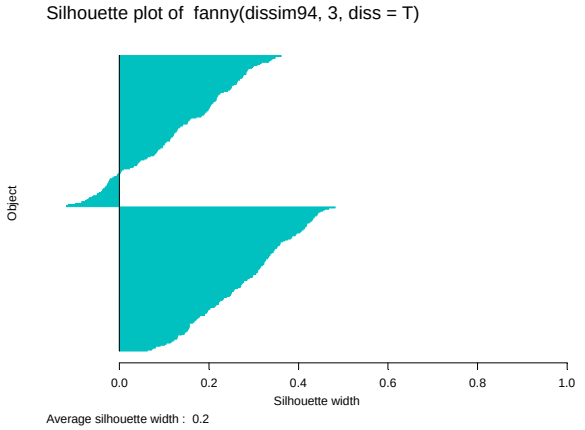


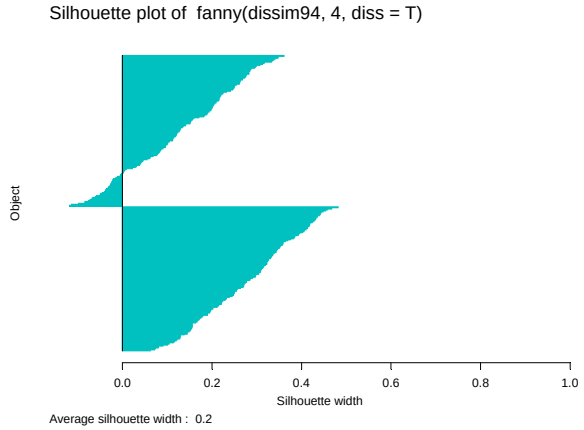


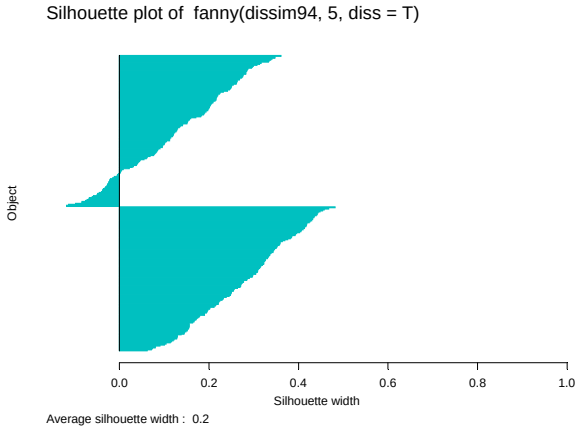

```
2 Cluster
dunn_coeff    normalized
0.5           3.552714e-015
3 Cluster
dunn_coeff    normalized
0.3333333    -3.330669e-016
4 Cluster
dunn_coeff    normalized
0.25         -1.850372e-016
5 Cluster
dunn_coeff    normalized
0.2           3.330669e-016
6 Cluster
dunn_coeff    normalized
0.1666667     0
```

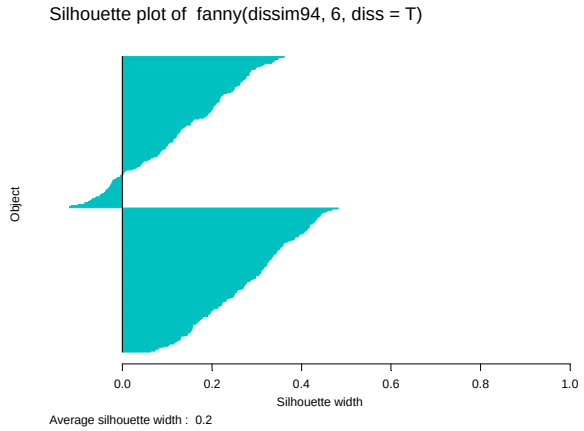
Abbildung B19b: Silhouettenplots für 2 bis 6 Clusterlösungen bei Fuzzy Clusterung, 1993












```
2 Cluster
dunn_coeff    normalized
0.5           2.220446e-015
3 Cluster
dunn_coeff    normalized
0.3333333     4.440892e-016
4 Cluster
dunn_coeff    normalized
0.25         -1.184238e-015
5 Cluster
dunn_coeff    normalized
0.2           3.885781e-015
6 Cluster
dunn_coeff     normalized
0.1666667     -2.442491e-016
```

Abbildung B19c: Silhouettenplots für 2 bis 6 Clusterlösungen bei Fuzzy Clusterung, 1994

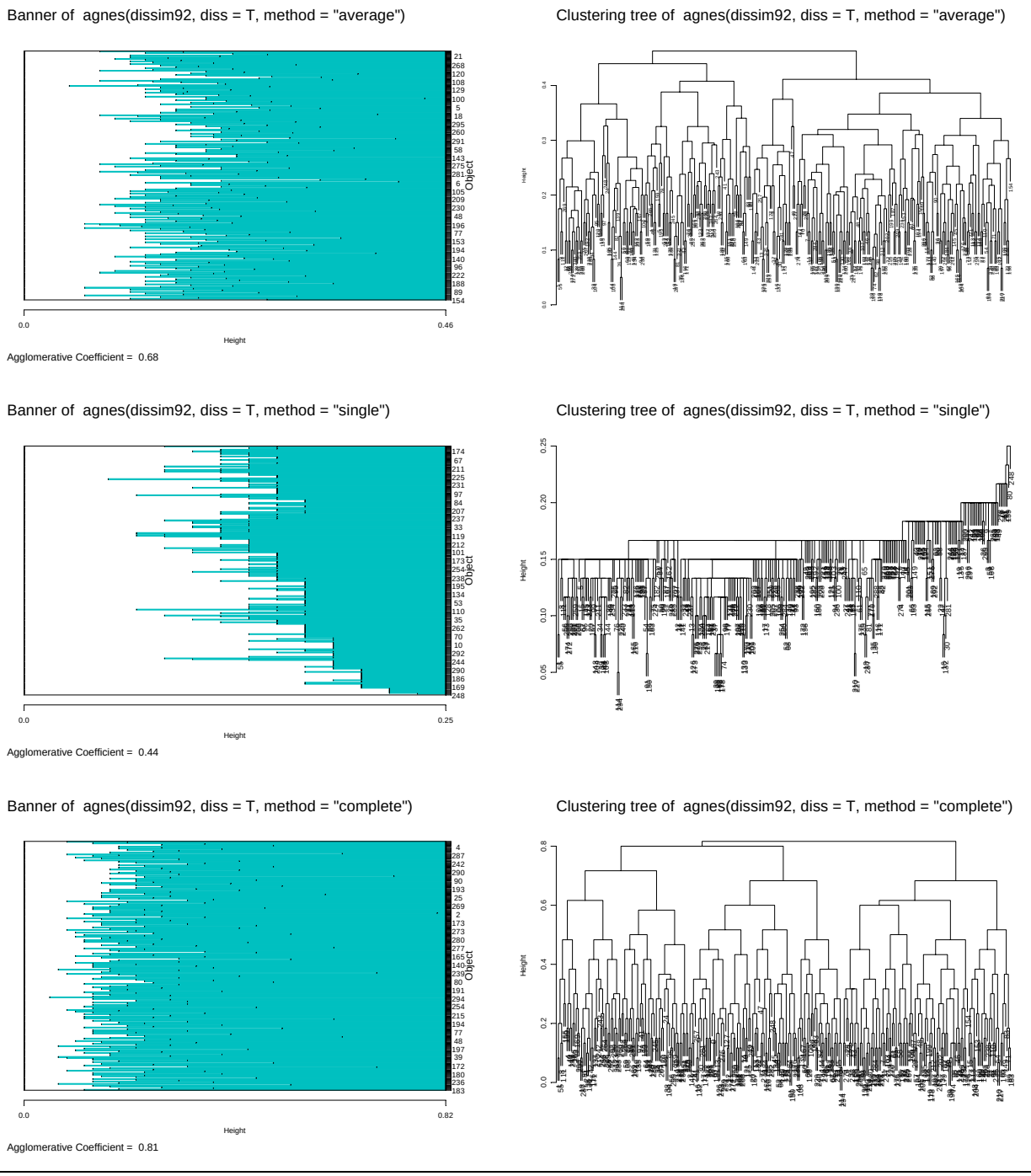


Abbildung B 20a: Bannerplots und Dendrogramme für hierarchische, agglomerative Clusteranalysen der Kennzahlenbetriebe, 1992

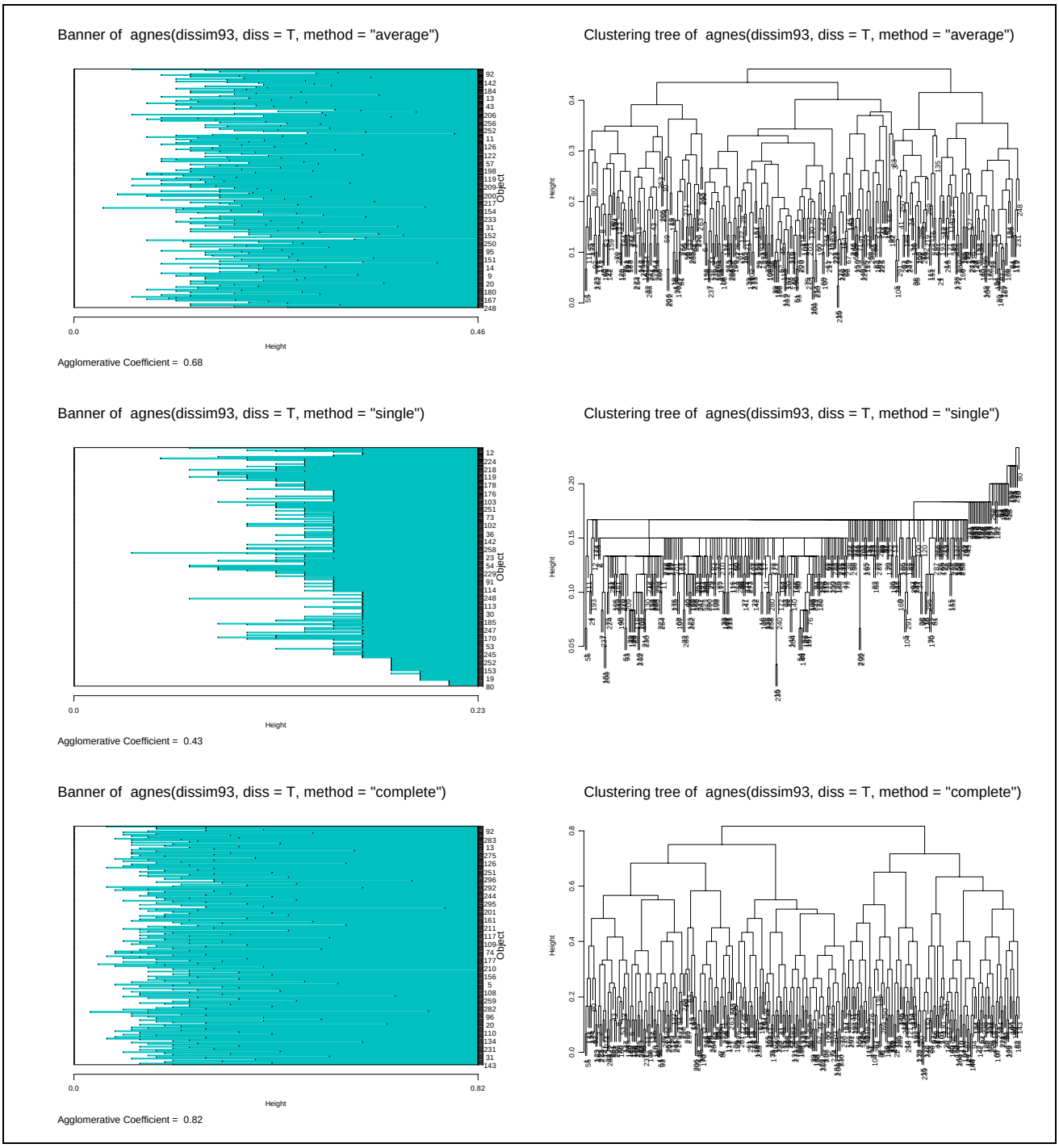


Abbildung B20b: Bannerplots und Dendrogramme für hierarchische, agglomerative Clusteranalysen der Kennzahlenbetriebe, 1993

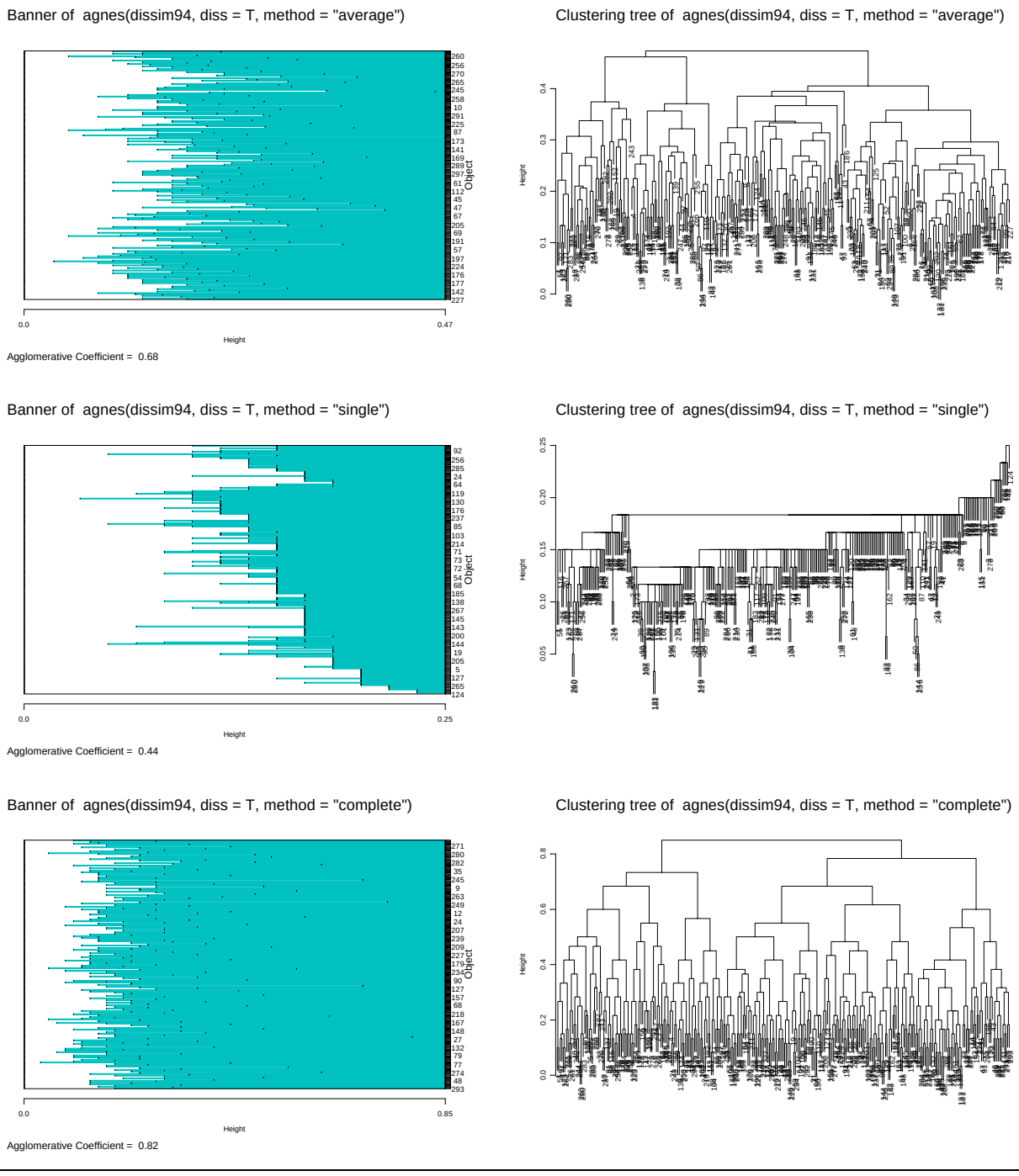


Abbildung B20c: Bannerplots und Dendrogramme für hierarchische, agglomerative Clusteranalysen der Kennzahlenbetriebe, 1994

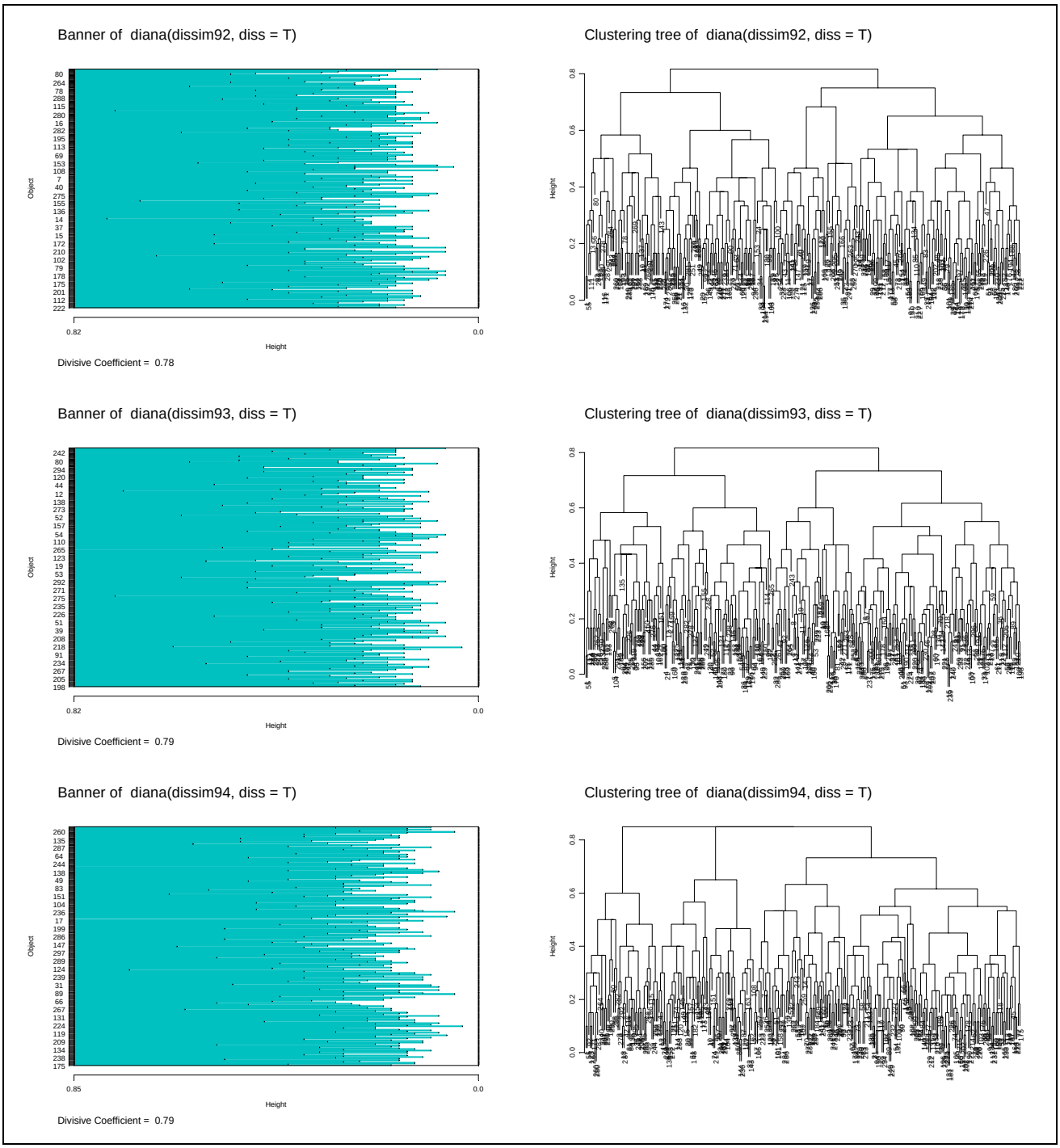
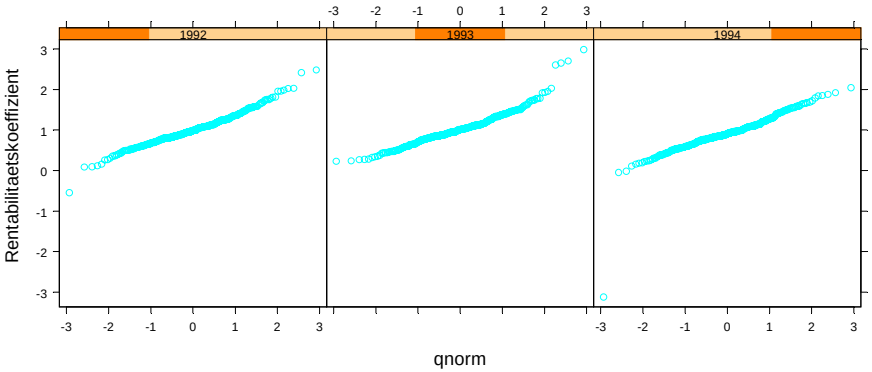


Abbildung B20d: Bannerplots und Dendrogramme für hierarchische, divisive Clusteranalyse der Kennzahlenbetriebe, 1992 bis 1994

a) Rentabilitätskoeffizient voller Datensatz



b) Rentabilitätskoeffizient, eingeschränkter Datensatz

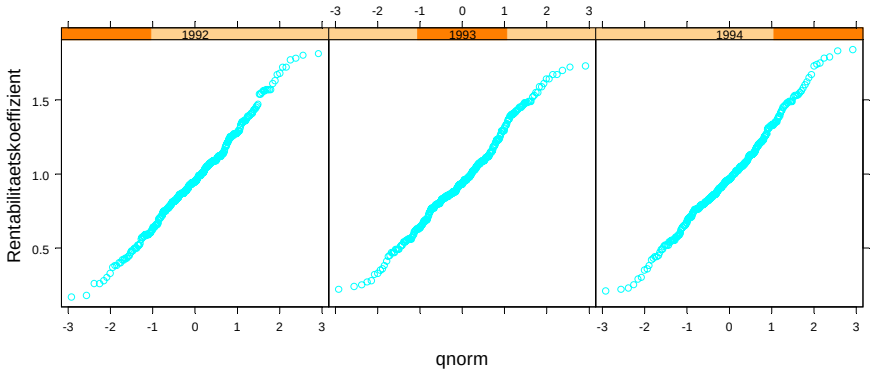


Abbildung B21: Normal-q-q-Plots für Kennzahl Rentabilitätskoeffizient im vollen (a)) und eingeschränkten (b)) Datensatz in 1992, 1993 und 1994

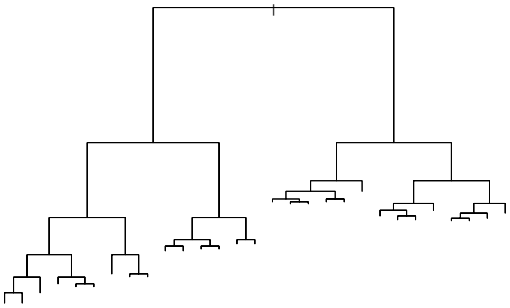
eingeschränkter Datensatz ohne die Kennzahlen der folgenden Objekte (siehe auch Übersicht B20):

145, 355, 472, 595, 634, 676, 799 in 1992

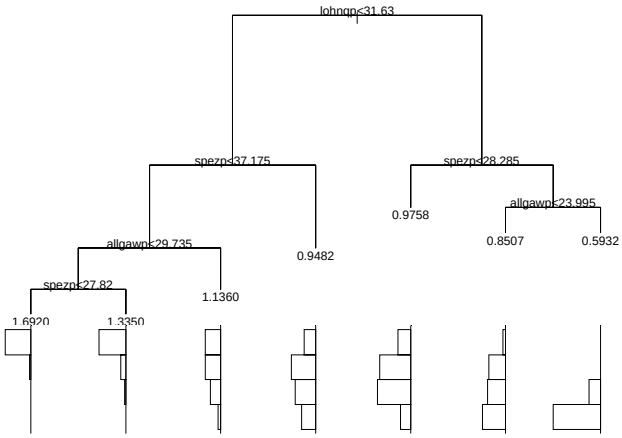
146, 326, 356, 416, 443, 533, 557, 611, 623, 656, 677, 788, 800, 881 in 1993

222, 297, 357, 393, 417, 486, 534, 636, 678 in 1994

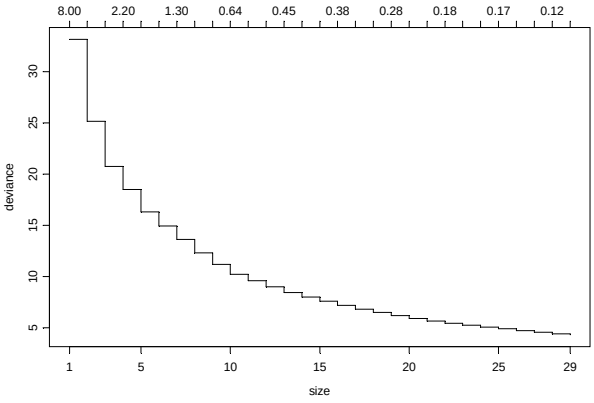
a) voller Regressionsbaum



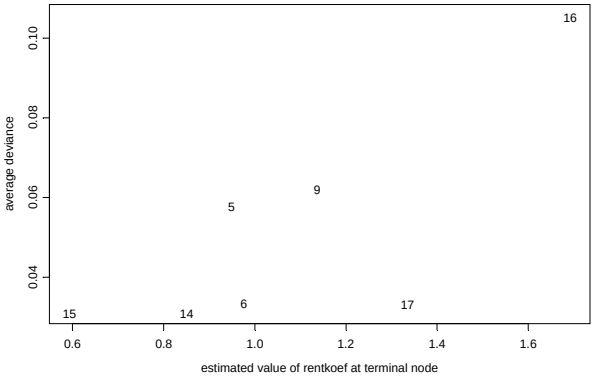
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



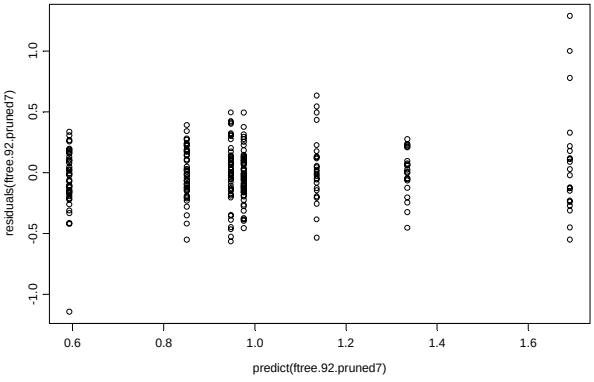
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

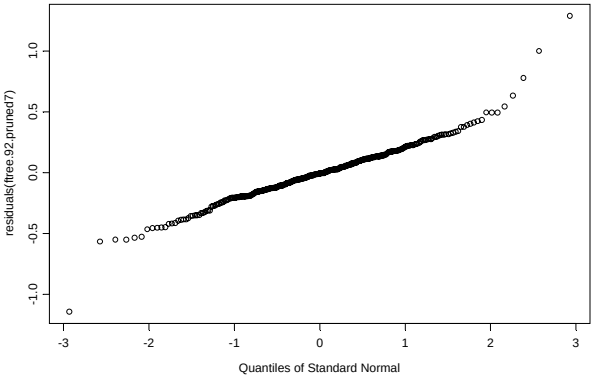
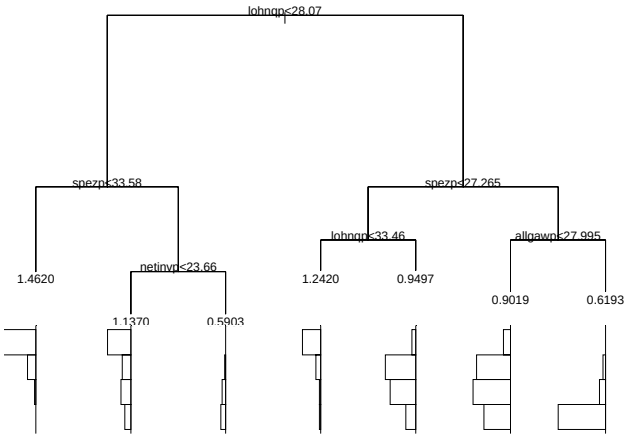


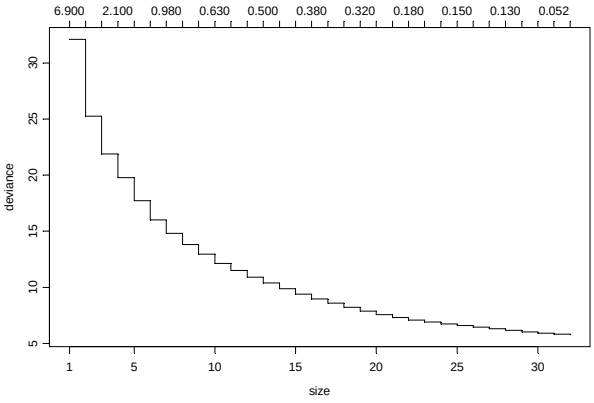
Abbildung B22: CART-Analyse 1992, abhängige Variable Rentabilitätskoeffizient, Verwendung der Gewichtung nach Ausreißertests

a) voller Regressionsbaum

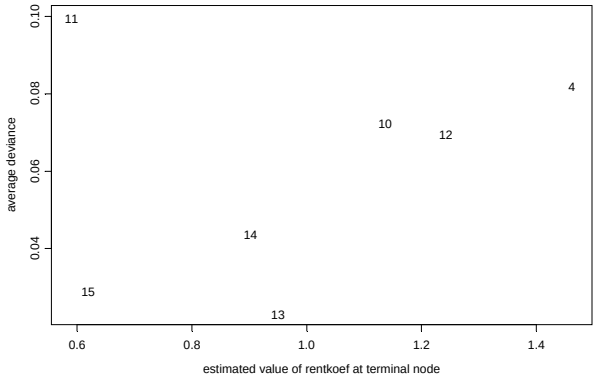
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



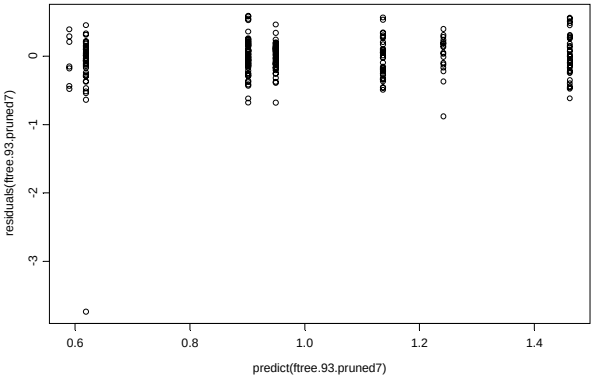
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

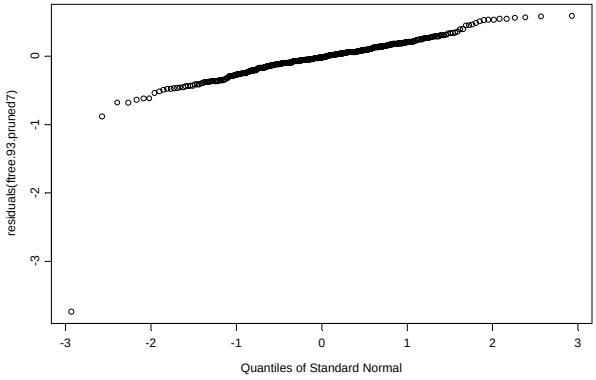
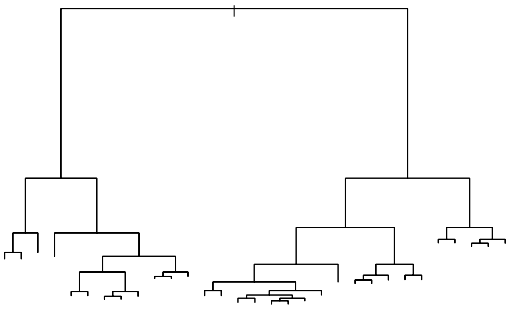
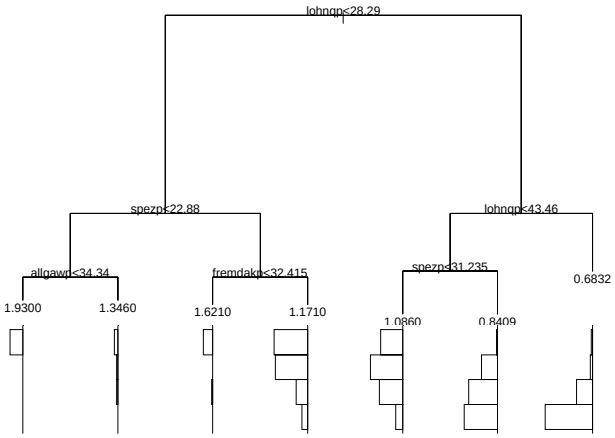


Abbildung B23: CART-Analyse 1993, abhängige Variable Rentabilitätskoeffizient, Verwendung der Gewichtung nach Ausreißertests

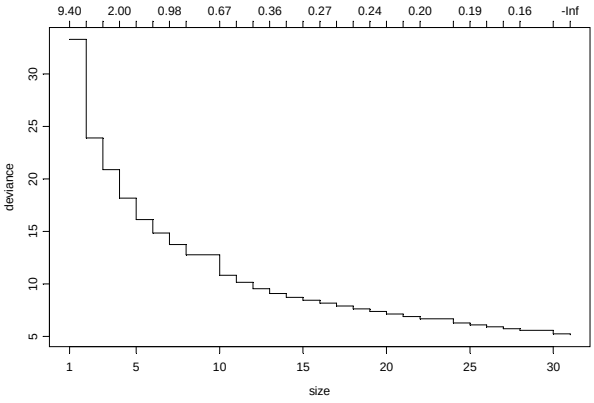
a) voller Regressionsbaum



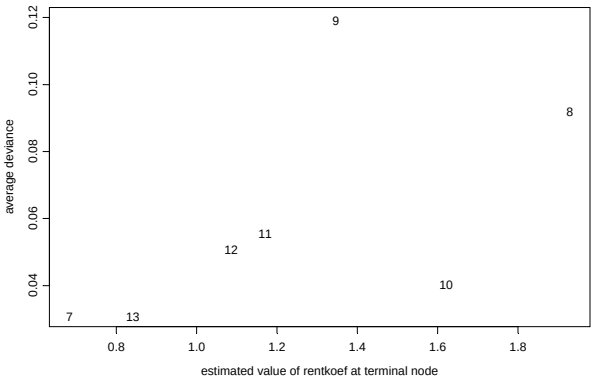
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



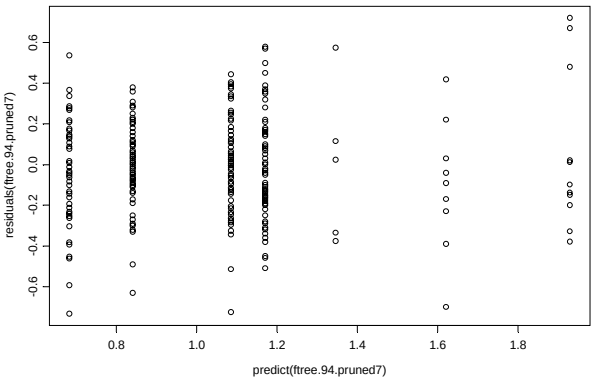
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

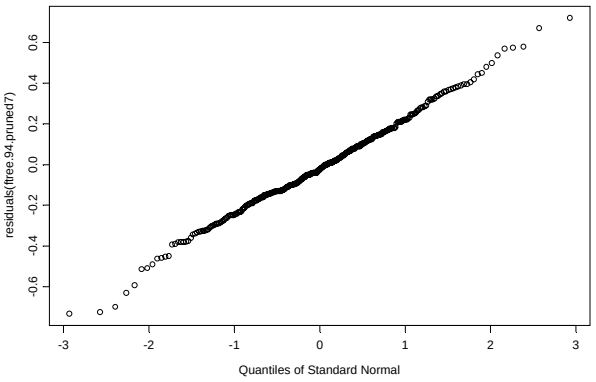
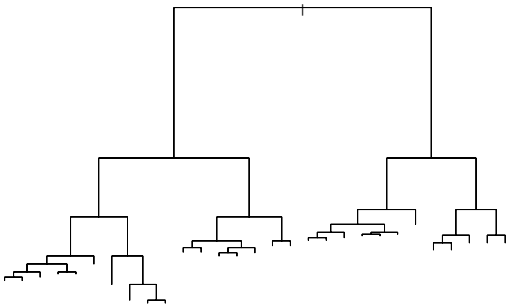
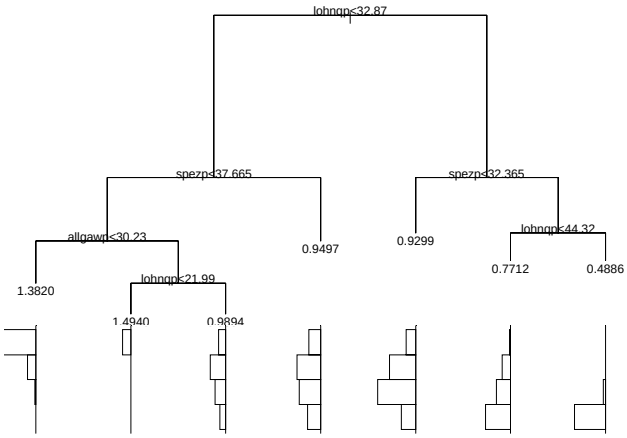


Abbildung B24: CART-Analyse 1994, abhängige Variable Rentabilitätskoeffizient, Verwendung der Gewichtung nach Ausreißertests

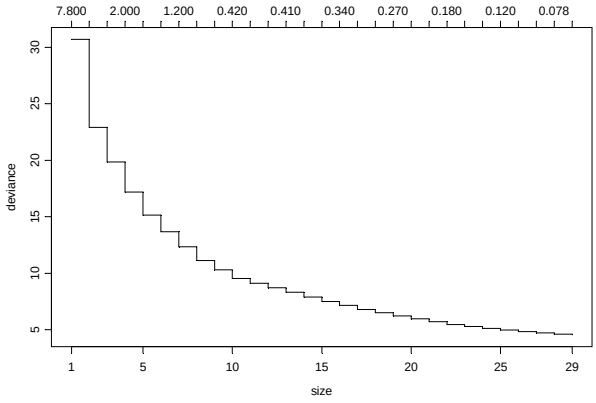
a) voller Regressionsbaum



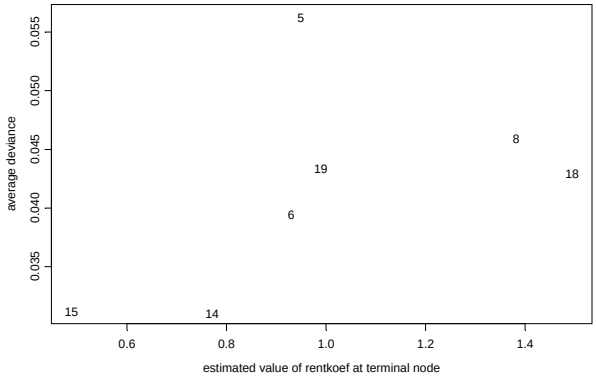
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



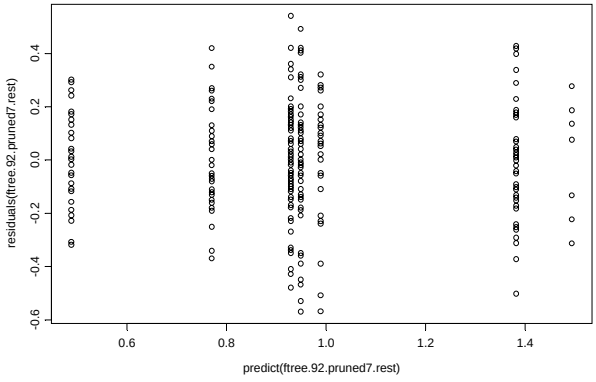
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

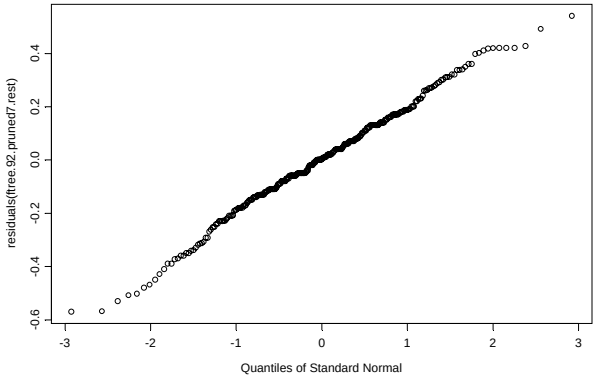
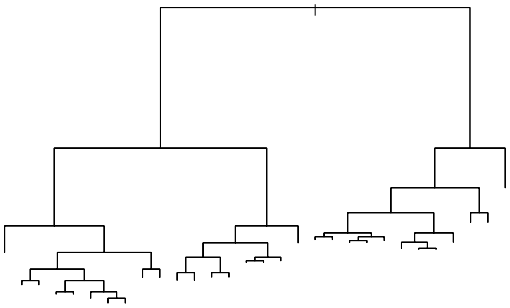
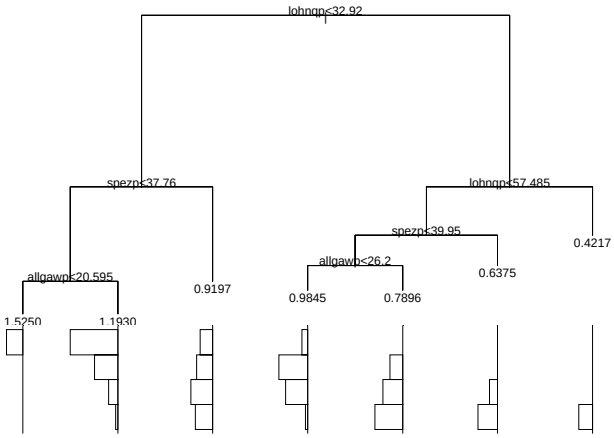


Abbildung B25: CART-Analyse 1992, abhängige Variable Rentabilitätskoeffizient, um Extremwerte verkleinerter Datensatz

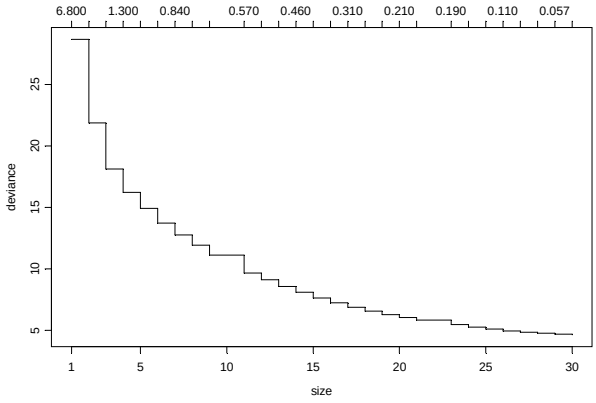
a) voller Regressionsbaum



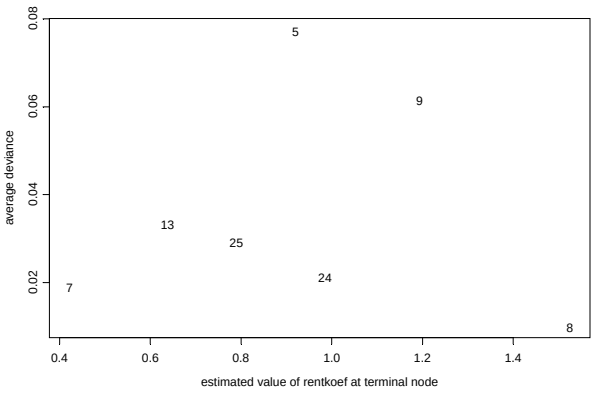
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



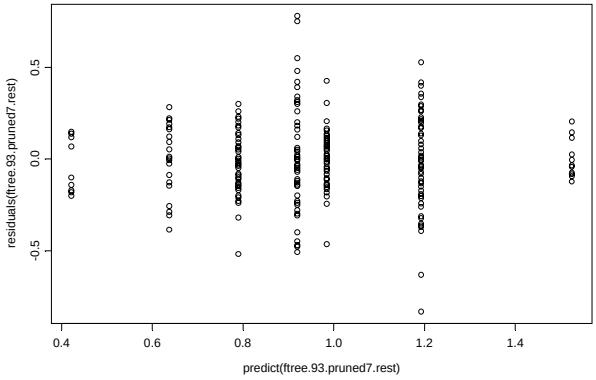
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

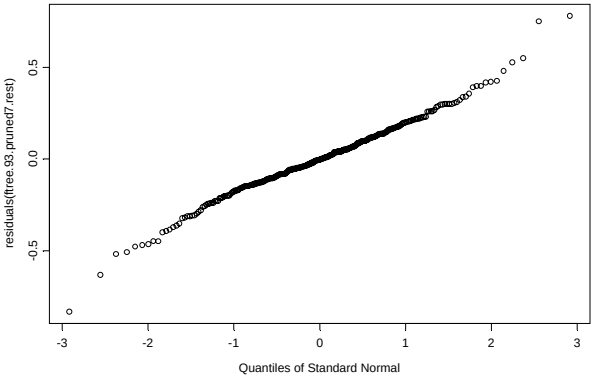
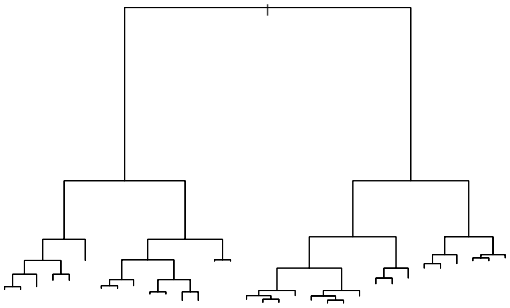
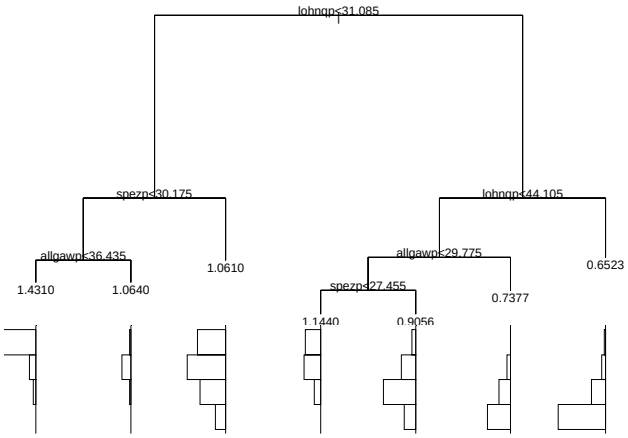


Abbildung B26: CART-Analyse 1993, abhängige Variable Rentabilitätskoeffizient, um Extremwerte verkleinerter Datensatz

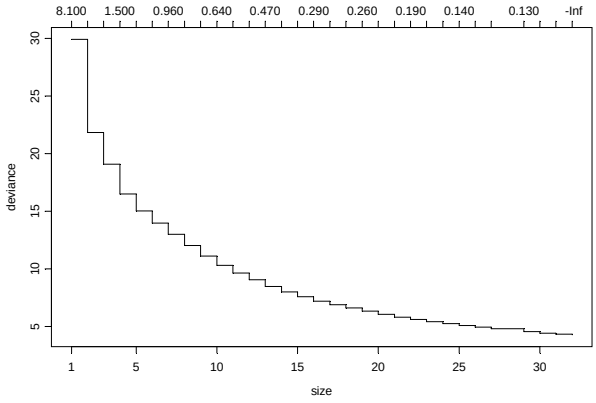
a) voller Regressionsbaum



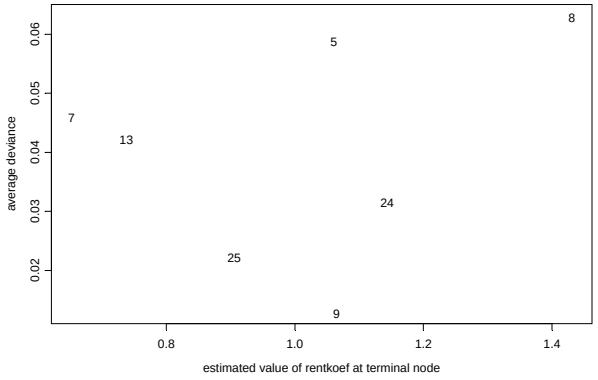
b) Regressionsbaum mit sieben Terminalknoten und Barcharts für Rentabilitätskoeffizent am jeweiligen Terminalknoten



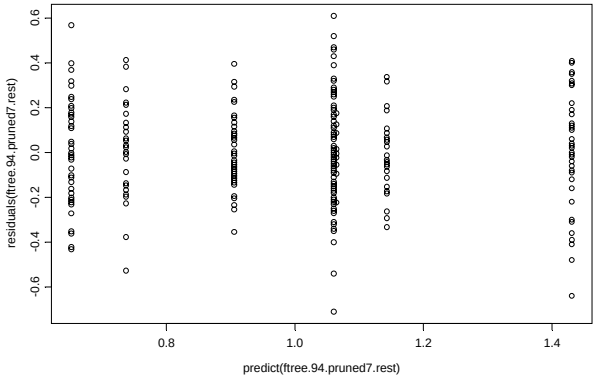
c) Cost complexity pruning



d) mittlere Residuendevianzen der Terminalknoten



e) Schätzwerte versus Residuen



f) Normal-q-q-Plot der Residuen

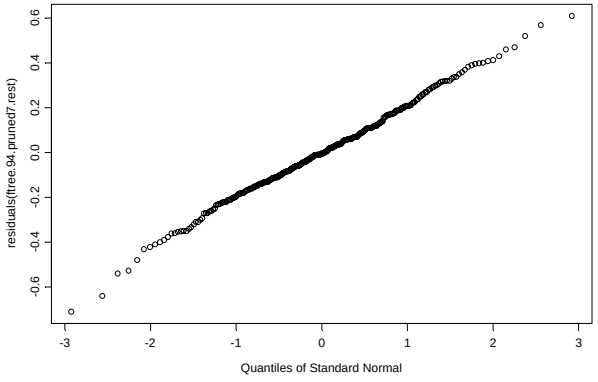


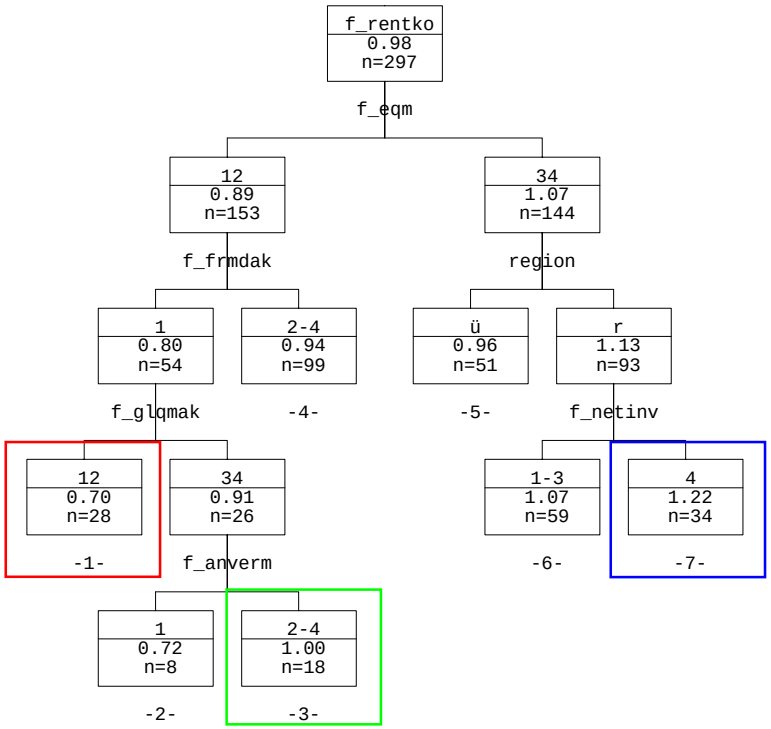
Abbildung B27: CART-Analyse 1994, abhängige Variable Rentabilitätskoeffizient, um Extremwerte verkleinerter Datensatz

Klassifikationsbaum 1992;

bestes Ergebnis,

schechtestes Ergebnis

weitere Splits möglich durch f_epertp mit p = 0.049



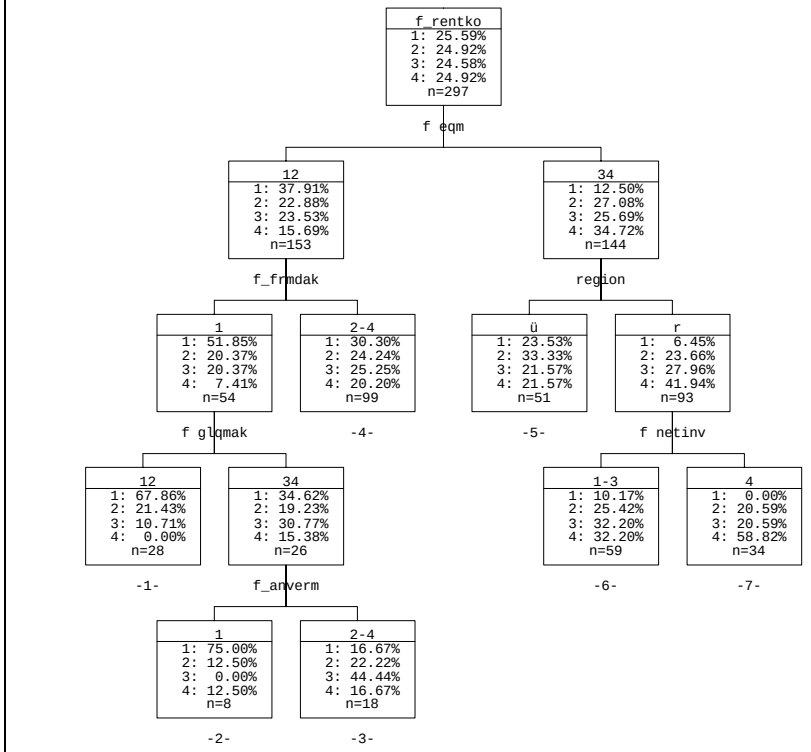


Abbildung B28: CHAID-Klassifikationsbaum; Analyse der ordinalskalierten Kennzahlen für 1992, abhängige Variable Rentabilitätskoeffizient

Klassifikationsbaum 1993;

bestes Ergebnis,

schechtestes Ergebnis

weitere Splits möglich durch f_glasqm mit p = 0.043 (Segment 3); f_eqm mit p = 0.0045 (Segemnt 4); f_eqm m,it p= 0.025 (Segment 5); f_epertp mit p = 0.015 (Segment 7); f_fkp mit p = 0.0061 (Segment 8)

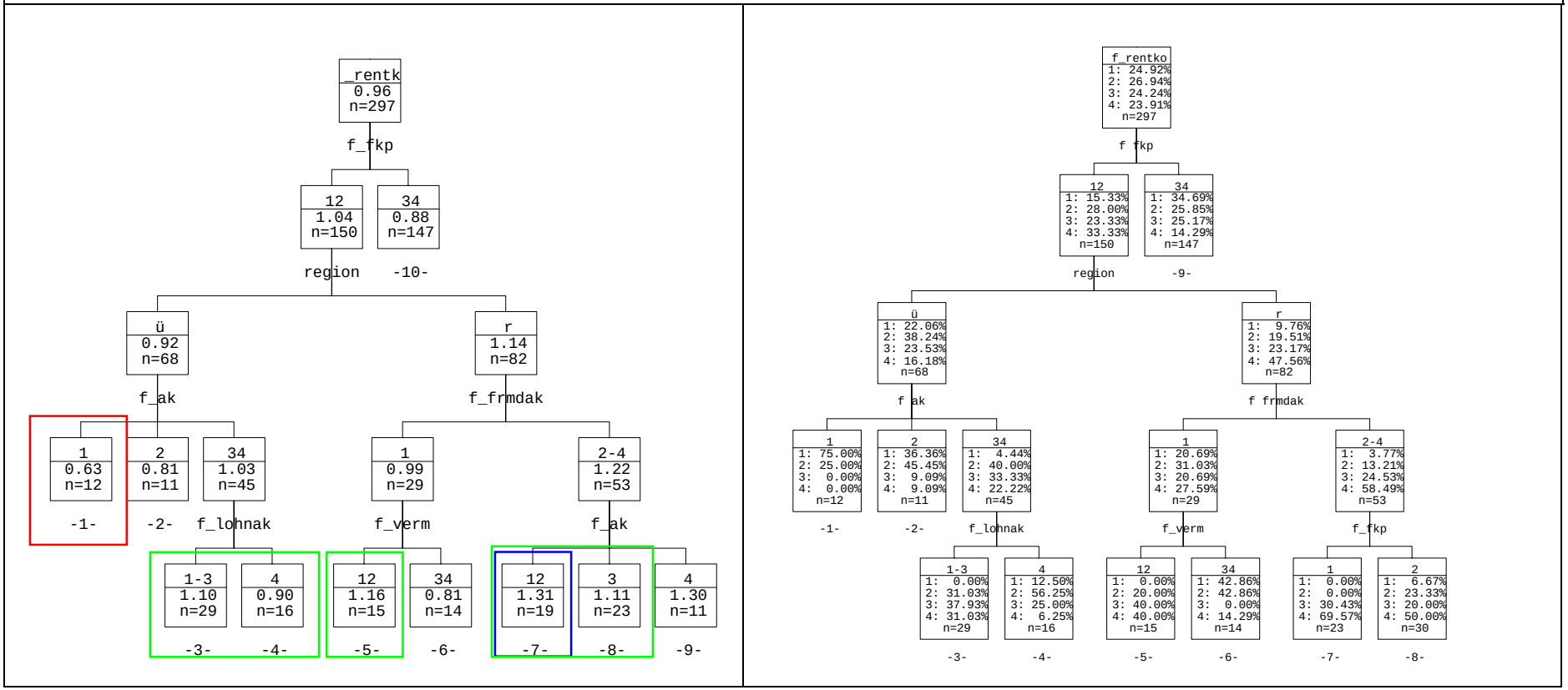


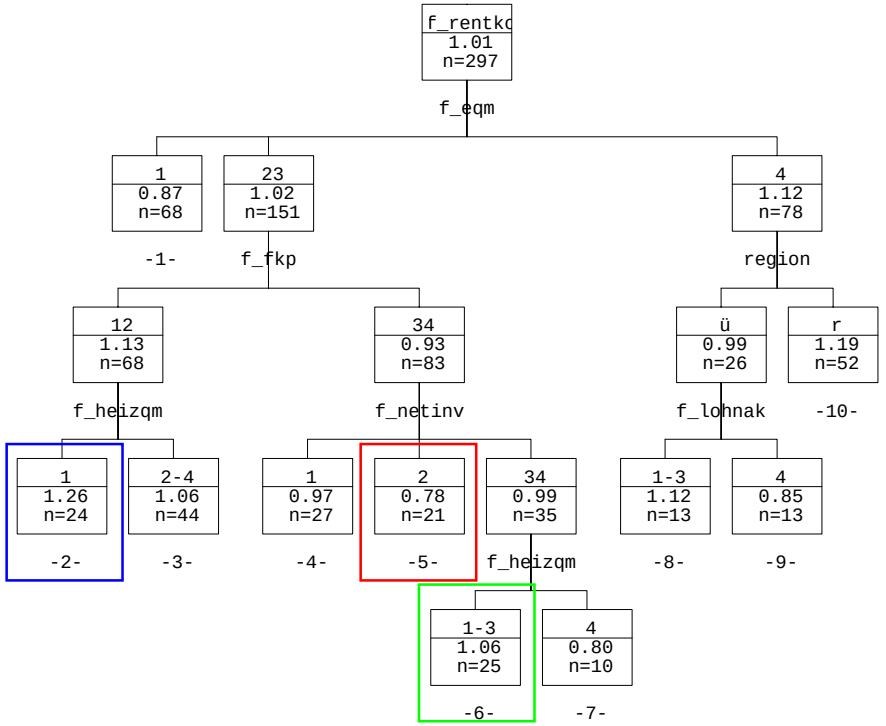
Abbildung B29: CHAID-Klassifikationsbaum; Analyse der ordinalskalierten Kennzahlen für 1993, abhängige Variable Rentabilitätskoeffizient

Klassifikationsbaum 1994;

bestes Ergebnis,

schechtestes Ergebnis

weitere Splits möglich durch f_fkp mit p = 0.012



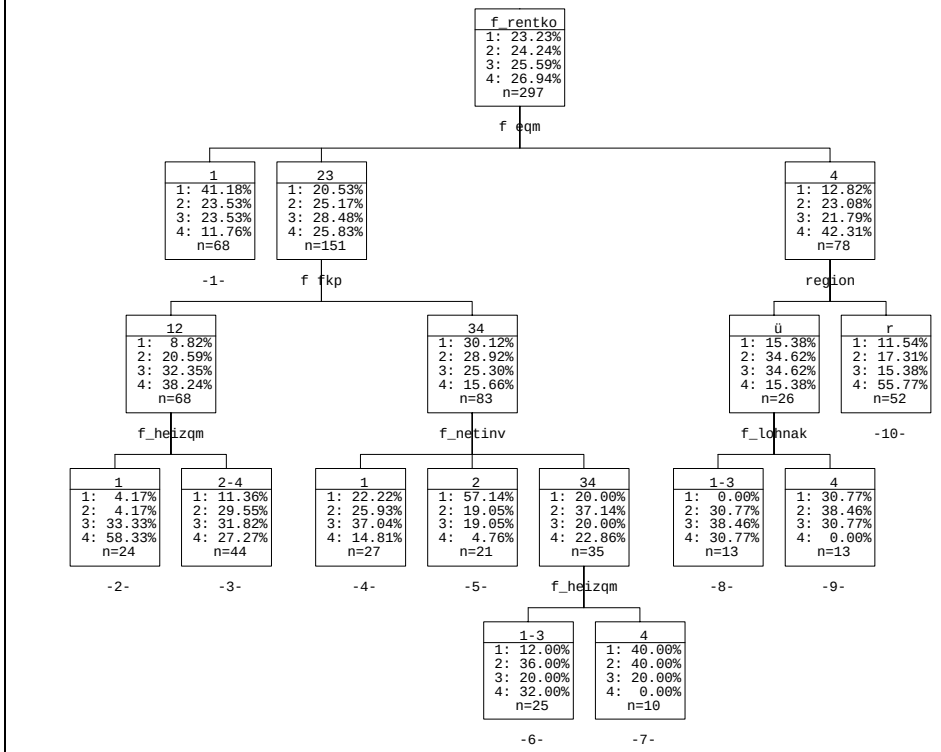


Abbildung B30: CHAID-Klassifikationsbaum; Analyse der ordinalskalierten Kennzahlen für 1994, abhängige Variable Rentabilitätskoeffizient

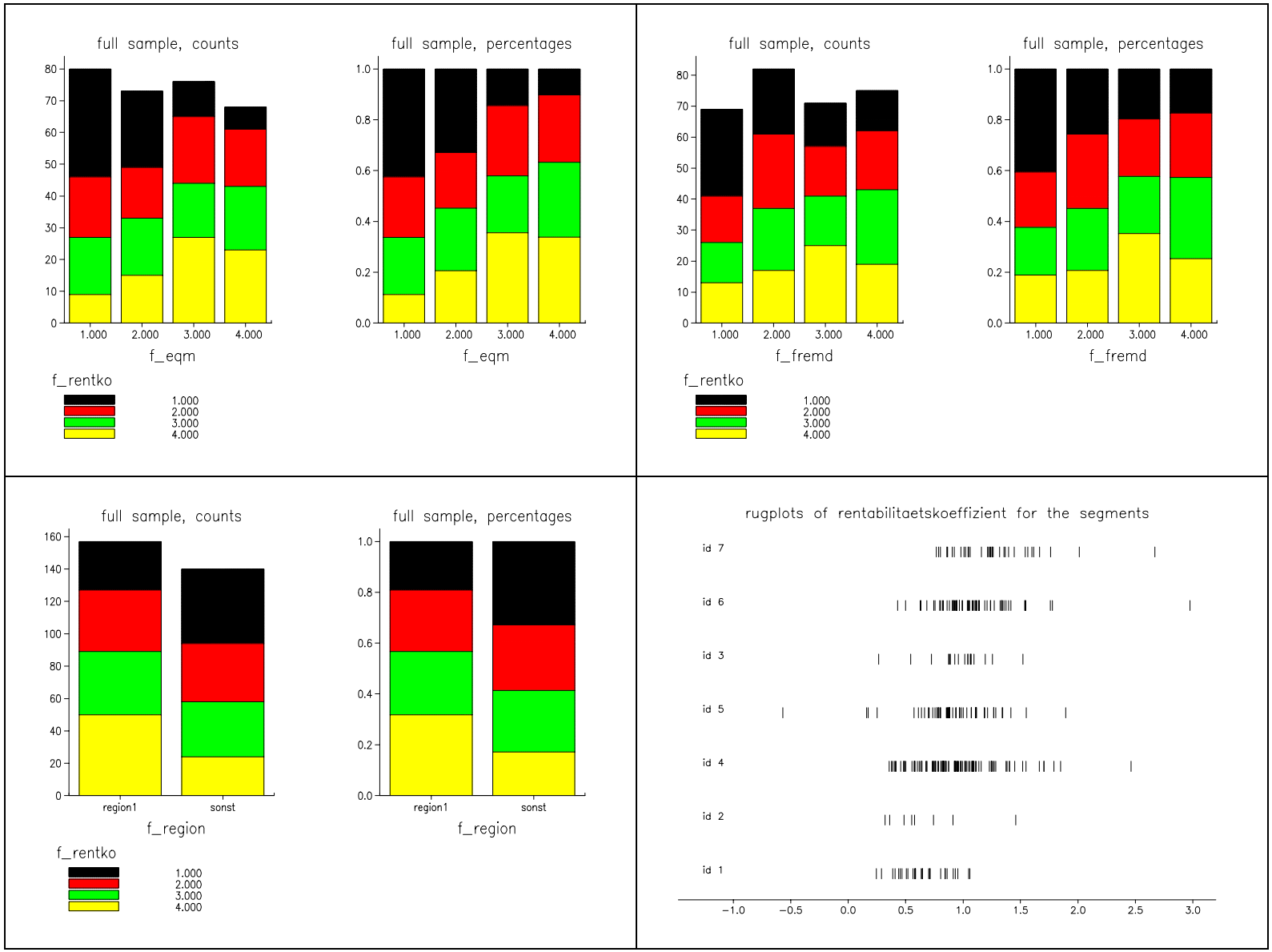


Abbildung B31: Balkendiagramme der wichtigsten Segmentierungsvariablen nach CHAID-Analyse für 1992 und Rugplot für die abhängige Variable in den Segmenten auf der untersten Ebene des Klassifikationsbaumes

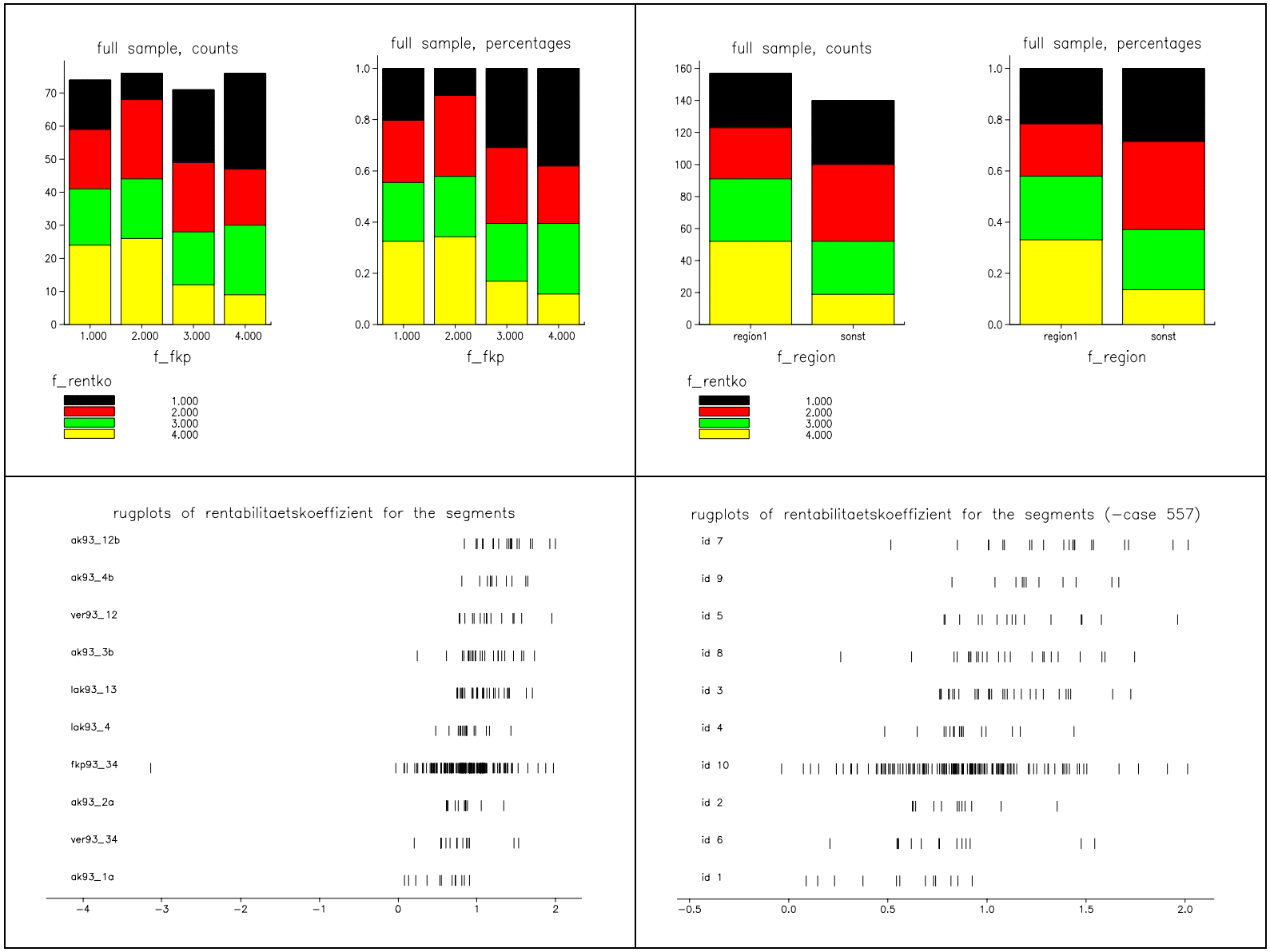


Abbildung B32: Balkendiagramme der wichtigsten Segmentierungsvariablen nach CHAID-Analyse für 1993 und Rugplots für die abhängige Variable in den Segmenten auf der untersten Ebene des Klassifikationsbaumes

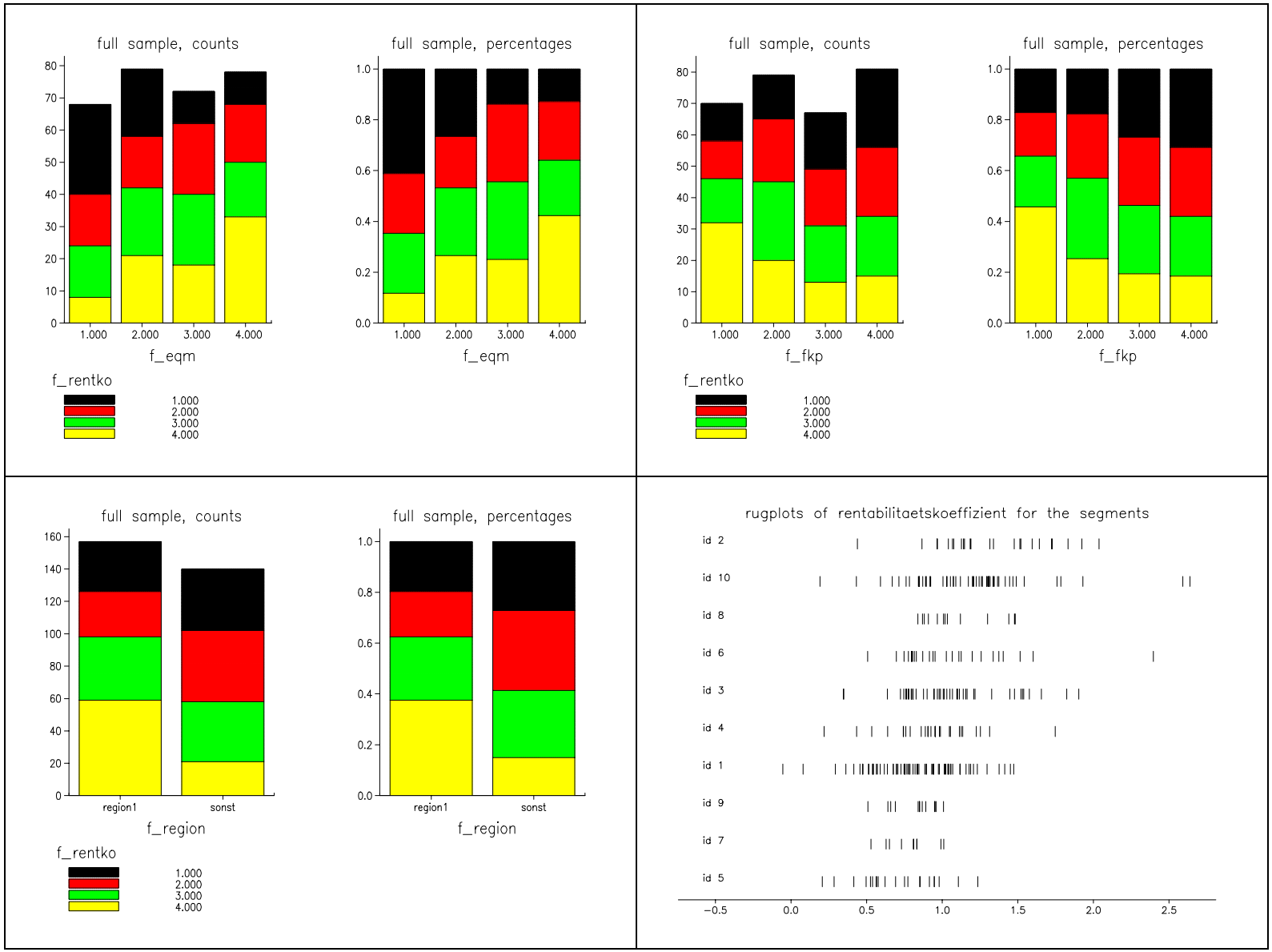


Abbildung B33: Balkendiagramme der wichtigsten Segmentierungsvariablen nach CHAID-Analyse für 1994 und Rugplot für die abhängige Variable in den Segmenten auf der untersten Ebene des Klassifikationsbaumes

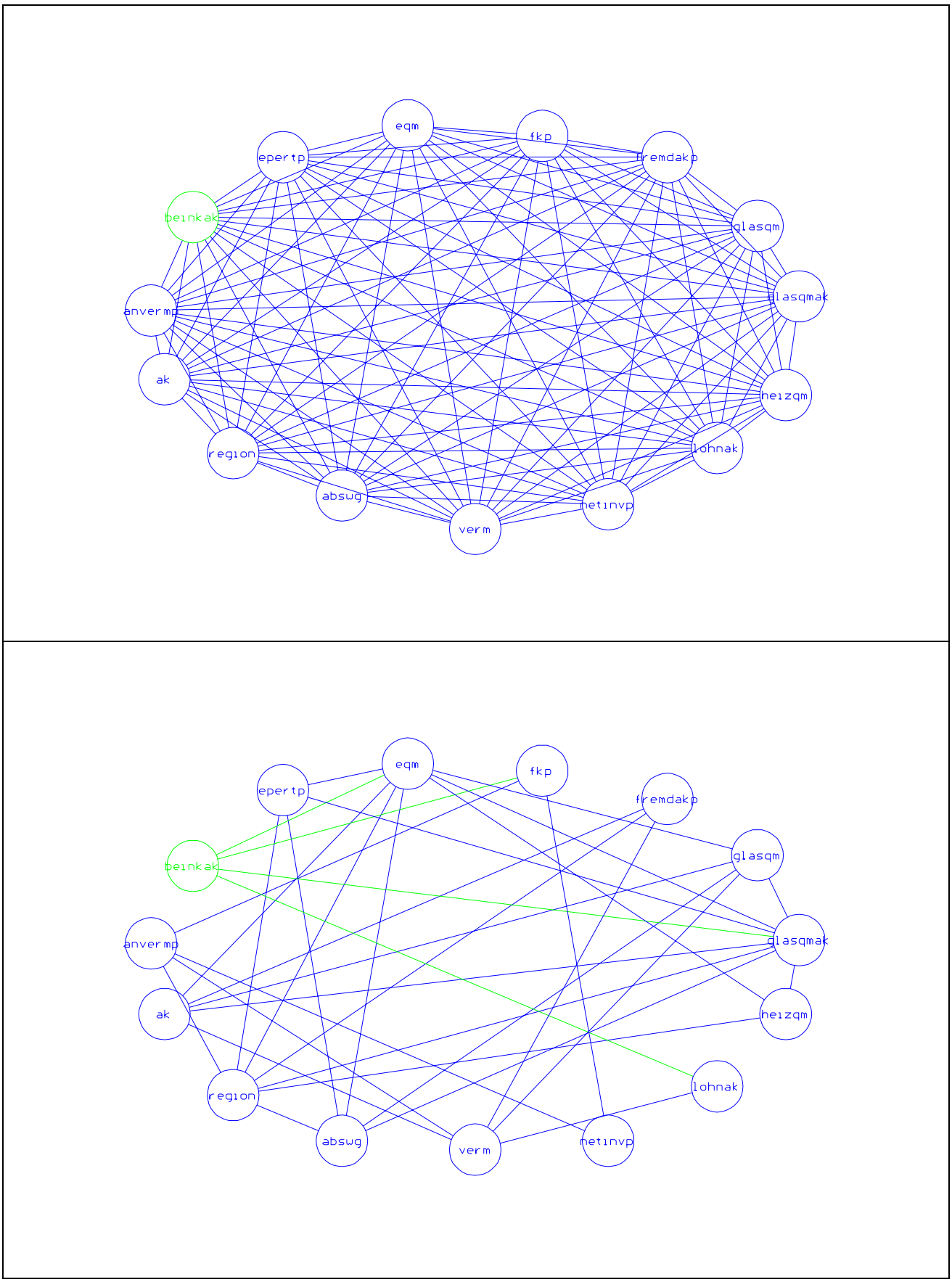
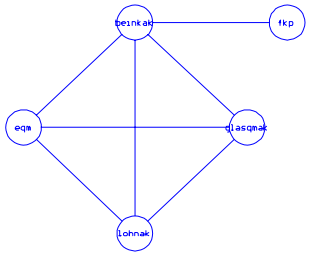


Abbildung B34: Beziehungsgeflecht eines vollständigen (oben) und eines auf direkte Beziehungen gescreenten (unten) graphischen Modells für die Analyse von 15 Kennzahlen im Jahr 1993; beteiligte Erfolgskennzahl: Betriebseinkommen/AK

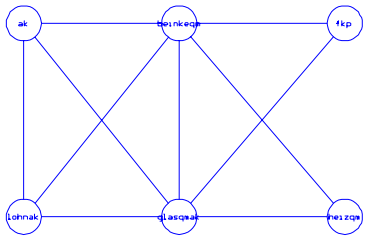
a) Graphisches Modell nach Screening und Rückwärts-Elimination für Betriebseinkommen/AK (beinkak), 1993



Deviance	df	p	original df	model
422.5	408	0.2996	756	(abde)(ac)

Degrees of freedom have been adjusted

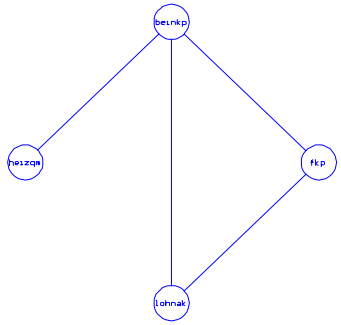
b) Graphisches Modell nach Screening und Rückwärts-Elimination für Betriebseinkommen/Eqm (beinkeqm), 1993



Deviance	df	p	original df	model
920.1	1612	1.0000	3744	(abdf)(acd)(ade)

Degrees of freedom have been adjusted

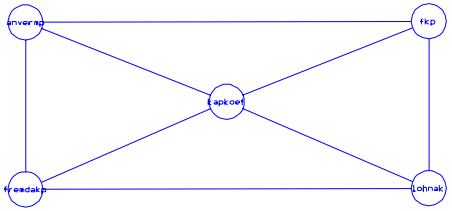
c) Graphisches Modell nach Screening und Rückwärts-Elimination für Betriebseinkommen in % BE (beinkp), 1993



Deviance	df	p	original df	model
173.3	177	0.5643	180	(abd)(ac)

Degrees of freedom have been adjusted

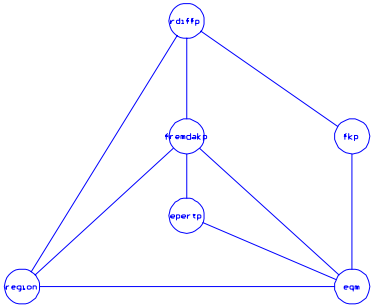
d) Graphisches Modell nach Screening und Rückwärts-Elimination für den Kapitalkoeffizienten (kapkoef), 1993



Deviance	df	p	original df	model
556.6	738	1.0000	828	(abc)(abd)(ace)(ade)

Degrees of freedom have been adjusted

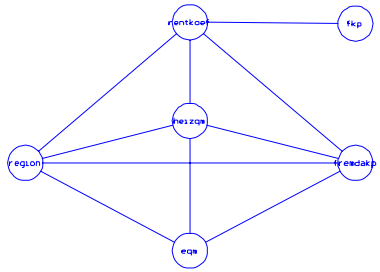
e) Graphisches Modell nach Screening und Rückwärts-Elimination für die Reinertragsdifferenz (rdiffp), 1993



Deviance	df	p	original df	model
872.5	1553	1.0000	1896	(abf)(aef)(bdf)(def)(cd)

Degrees of freedom have been adjusted

f) Graphisches Modell nach Screening und Rückwärts-Elimination für den Rentabilitätskoeffizienten (rentkoef), 1993

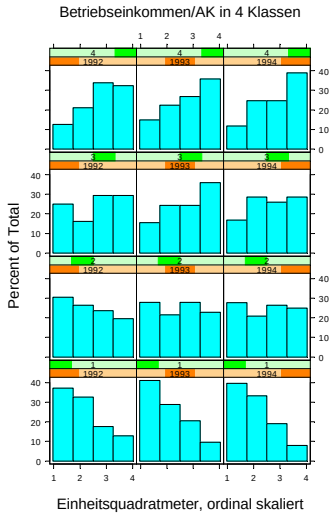


Deviance	df	p	original df	model
682.2	925	1.0000	1812	(abef)(bcef)(ad)

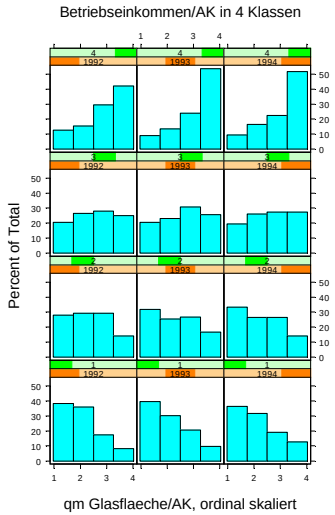
Degrees of freedom have been adjusted

Abbildung B35: Graphische Modelle nach Rückwärts-Elimination 1993

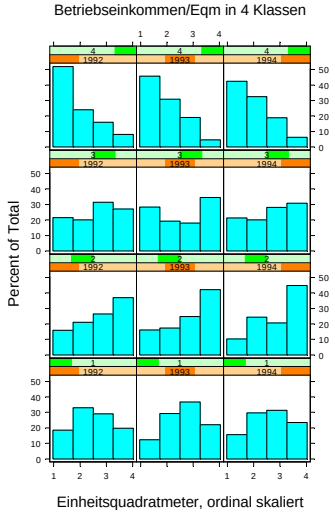
a) Betriebseinkommen/AK und Einheitsquadratmeter



b) Betriebseinkommen/AK und qm Glasfläche/AK



c) Betriebseinkommen/Eqm und Einheitsquadratmeter



d) Betriebseinkommen/Eqm und qm Glasfläche/AK

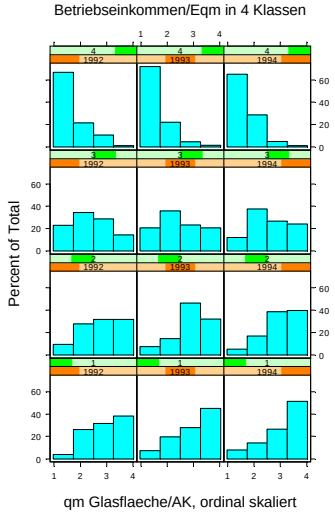
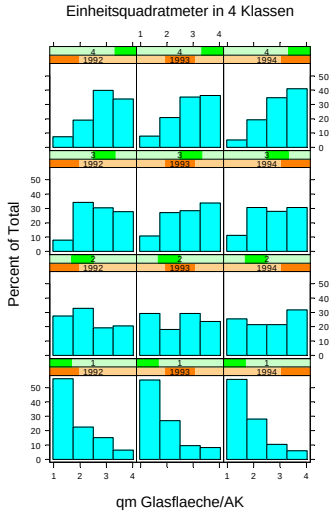
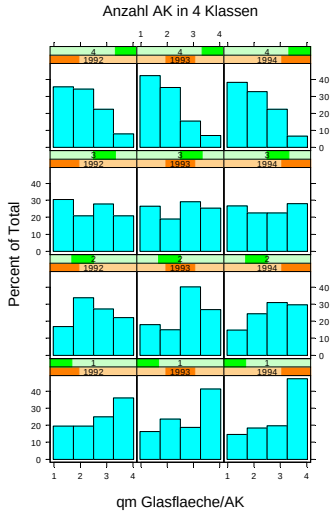


Abbildung B36: Beziehungen von Betriebseinkommen/AK und Betriebseinkommen/Eqm zu Einheitsquadratmeter beziehungsweise qm Glasfläche/AK, 1992 bis 1994

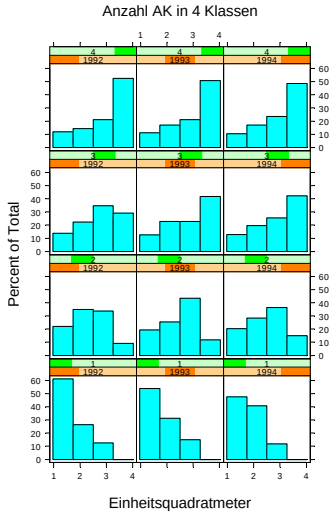
a) Einheitsquadratmeter und qm Glasfläche/AK



b) Anzahl AK und qm Glasfläche/AK



c) Anzahl AK und Einheitsquadratmeter



d) Anzahl AK, Einheitsquadratmeter und qm Glasfläche/AK, 1993

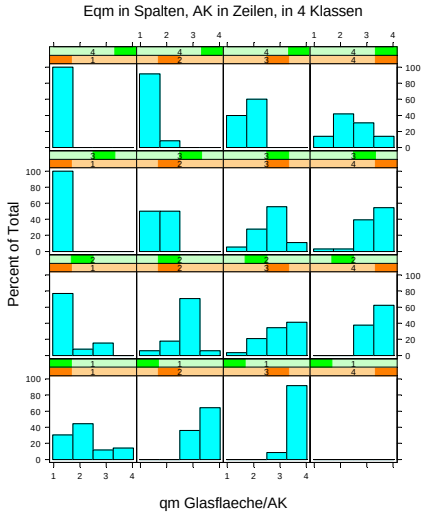
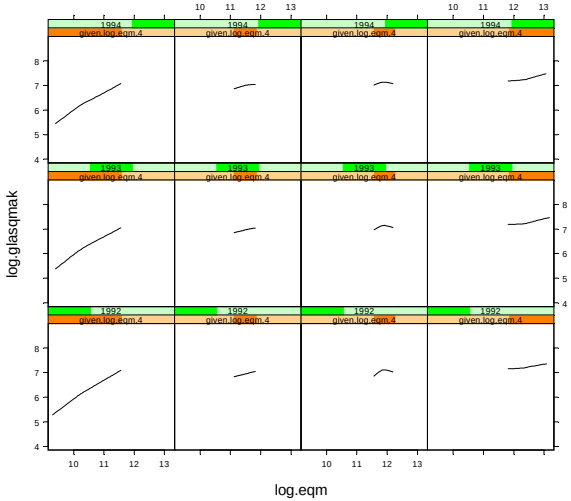
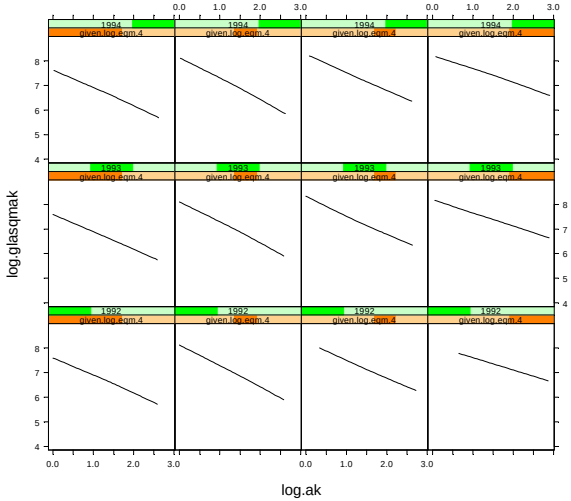


Abbildung B37: Beziehungen von Einheitsquadratmeter, Anzahl AK und Glasläche/AK, 1992 bis 1994

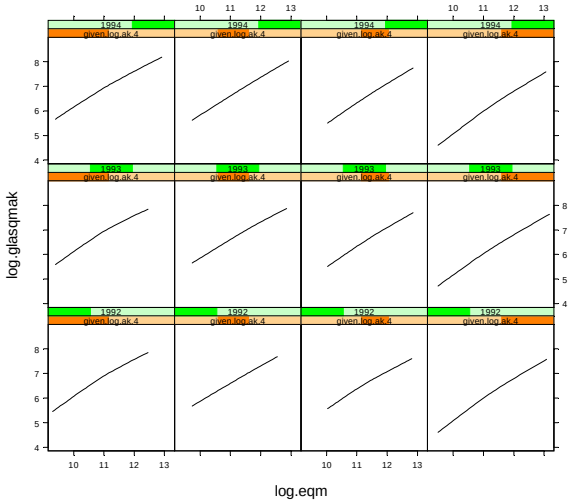
a) Spalten log(eqm), Panels log(glasqmak) versus log(eqm)



b) Spalten log(eqm), Panels log(glasqmak) versus log(ak)



c) Spalten log(ak), Panels log(glasqmak) versus log(eqm)



d) Spalten log(ak), Panels log(glasqmak) versus log(ak)

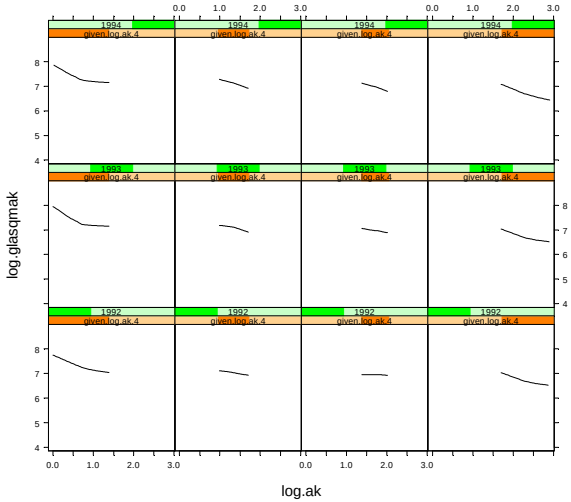
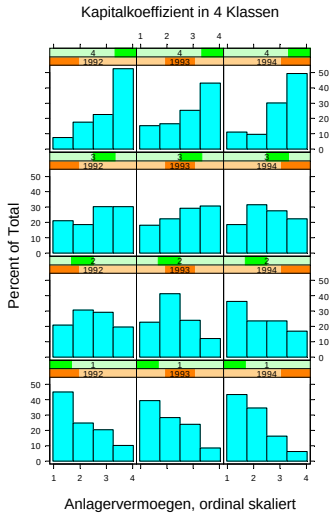
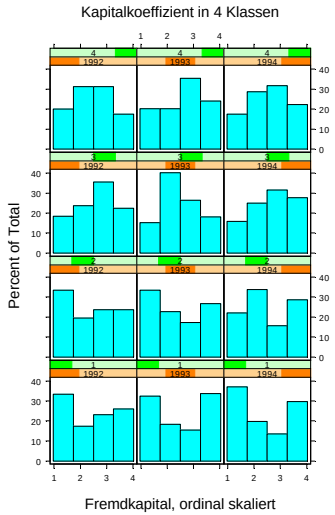


Abbildung B38: Beziehungen von Einheitsquadratmeter, Anzahl AK und qm Glasfläche/AK; Loess-Regressionslinien der log-transformierten Variablen in den Panels mit 50% überlappenden Intervallen, 1992 bis 1994

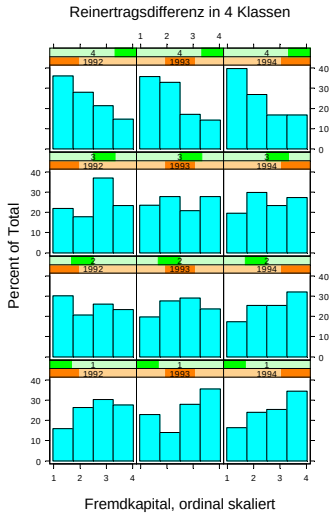
a) Kapitalkoeffizient und Anlagevermögen



b) Kapitalkoeffizient und Fremdkapital



c) Reinertragsdifferenz und Fremdkapital



d) Rentabilitätskoeffizient und Fremdkapital

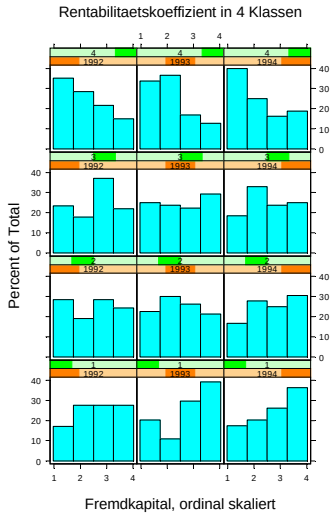
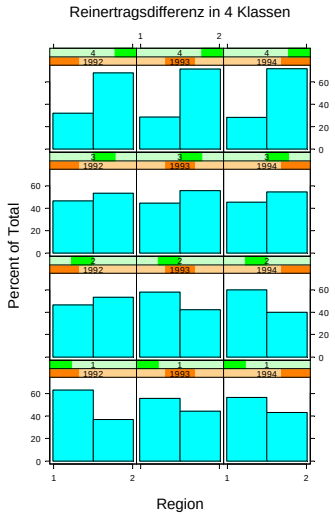
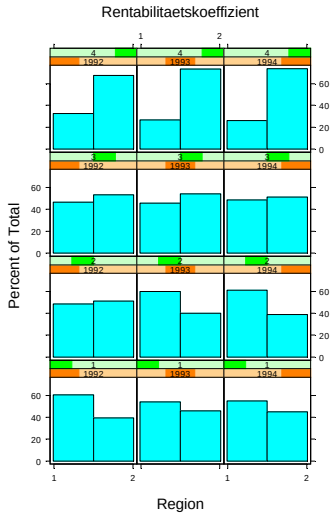


Abbildung B39: Beziehungen von Fremdkapital und Anlagevermögen zu Kapitalkoeffizient, Reinertragsdifferenz und Rentabilitätskoeffizient, 1992 bis 1994

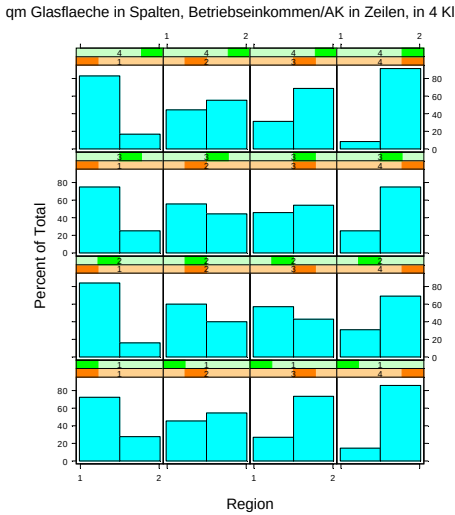
a) Reinertragsdifferenz und Region



b) Rentabilitätskoeffizient und Region



c) Betriebseinkommen/AK, Region und Glasfläche/AK 1993



d) Rentabilitätskoeffizient, Region und Glasfläche/AK 1993

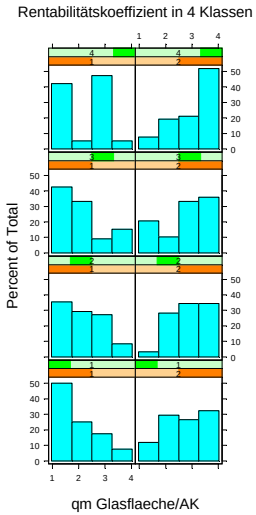
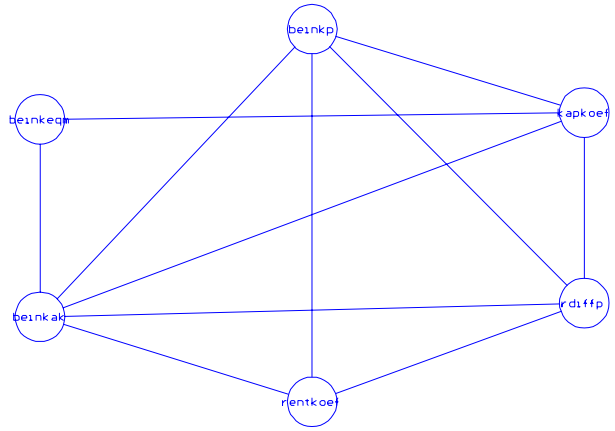
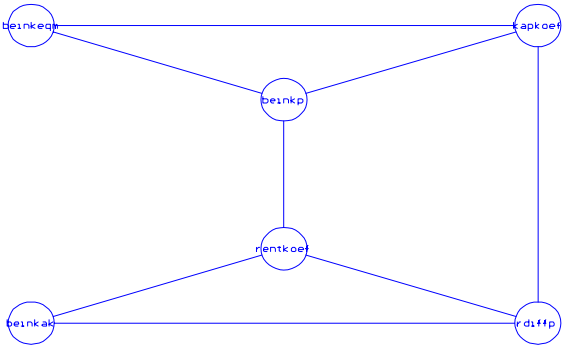


Abbildung B40: Beziehungen von Region und qm Glasfläche/AK zu Reinertragsdifferenz, Rentabilitätskoeffizient und Bertriebseinkommen/AK, 1992 bis 1994

a) 1992



b) 1993



c) 1994

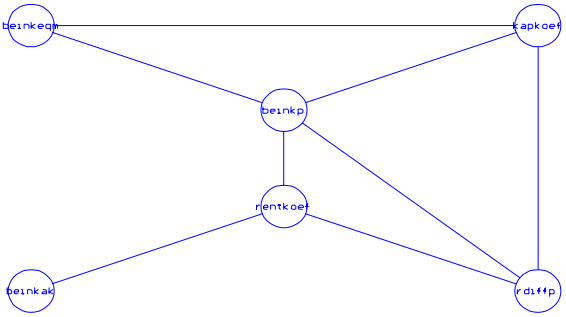
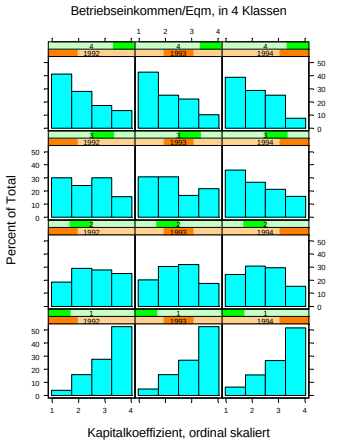
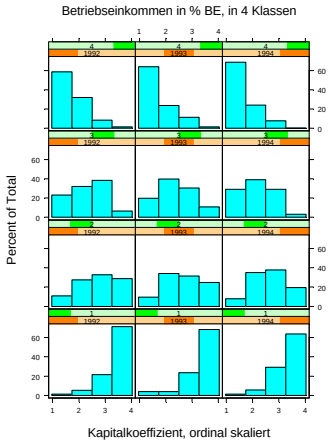


Abbildung B41: Graphische Modelle für sechs Erfolgskennzahlen nach Rückwärts-Elimination

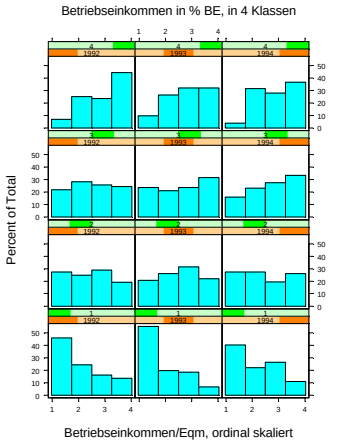
a) Betriebseinkommen/Eqm und Kapitalkoeffizient



b) Betriebseinkommen in % BE und Kapitalkoeffizient



c) Betriebseinkommen in % BE und Betriebseinkommen/Eqm



d) Betriebseinkommen in % BE und Rentabilitätskoeffizient

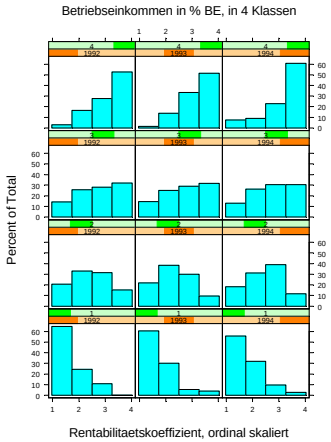
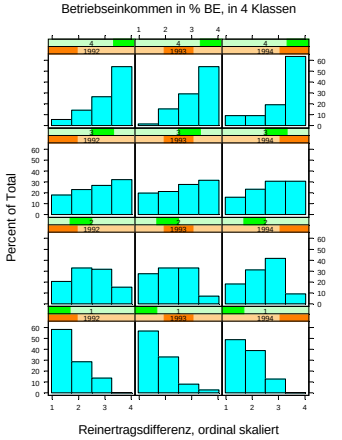
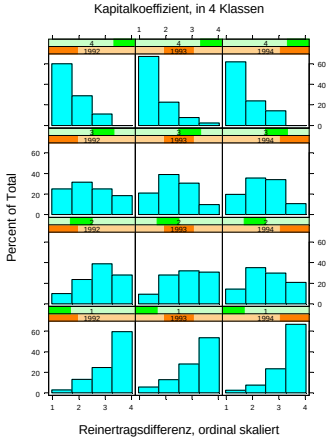


Abbildung B42: Beziehungen von Betriebseinkommen/Eqm, Betriebseinkomen in % BE, Kapitalkoeffizient und Rentabilitätskoeffizient, 1992 bis 1994

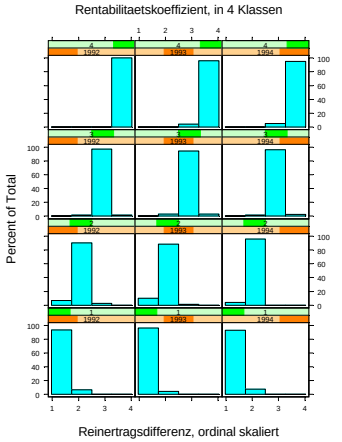
a) Betriebseinkommen in % BE und Reinertragsdifferenz



b) Kapitalkoeffizient und Reinertragsdifferenz



c) Rentabilitätskoeffizient und Reinertragsdifferenz



d) Betriebseinkommen/AK und Rentabilitätskoeffizient

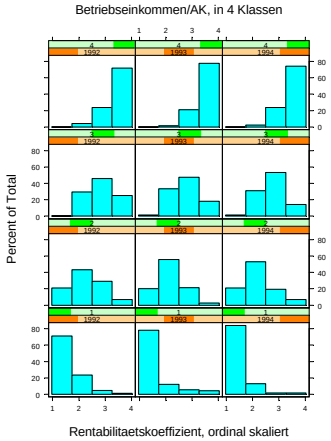


Abbildung B43: Beziehungen von Betriebseinkommen/Eqm, Betriebseinkommen in % BE, Kapitalkoeffizient und Rentabilitätskoeffizient, 1992 bis 1994

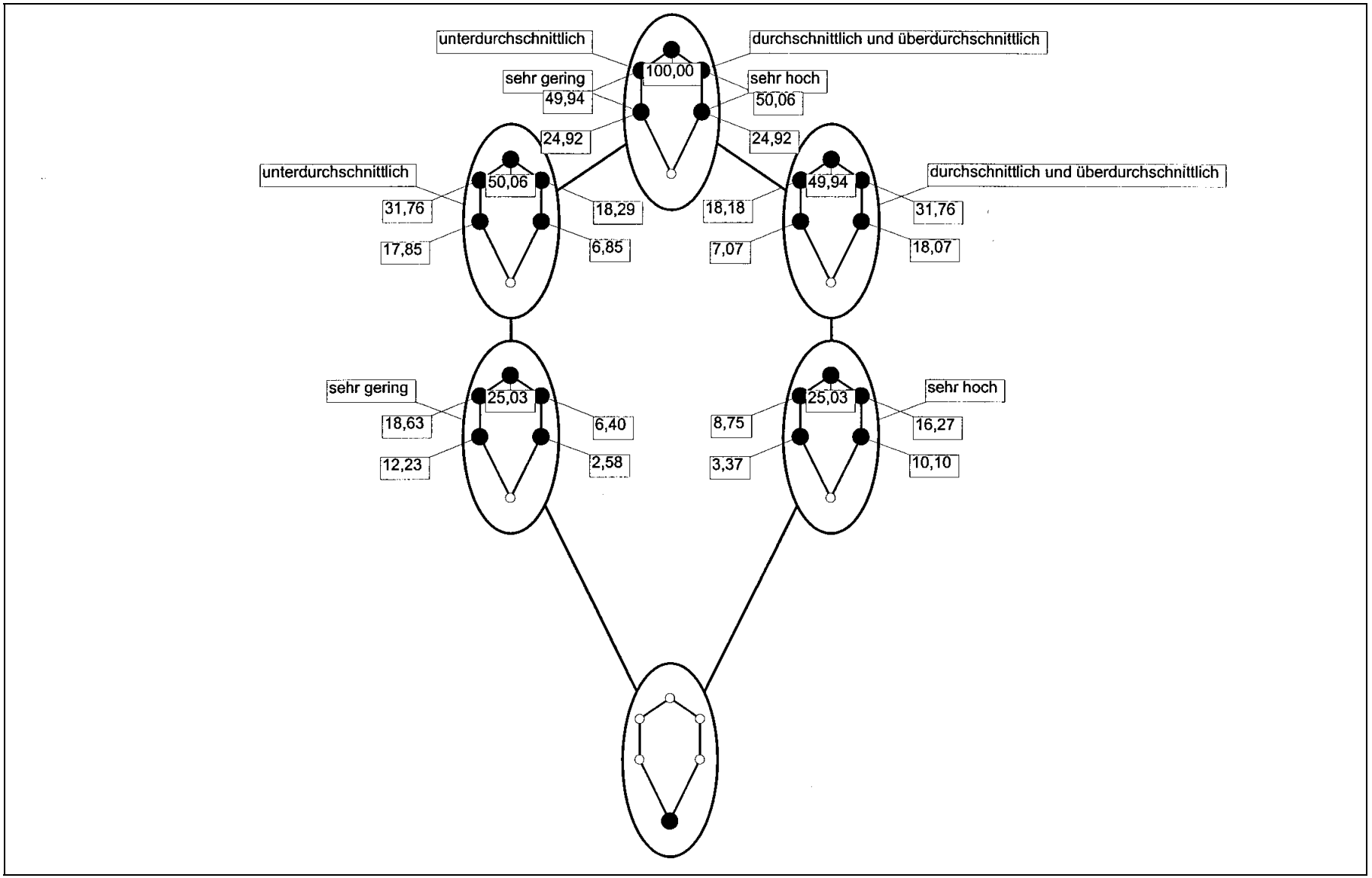


Abbildung B44: Liniendiagramm für Betriebseinkommen je AK und Lohn je entlohnte AK

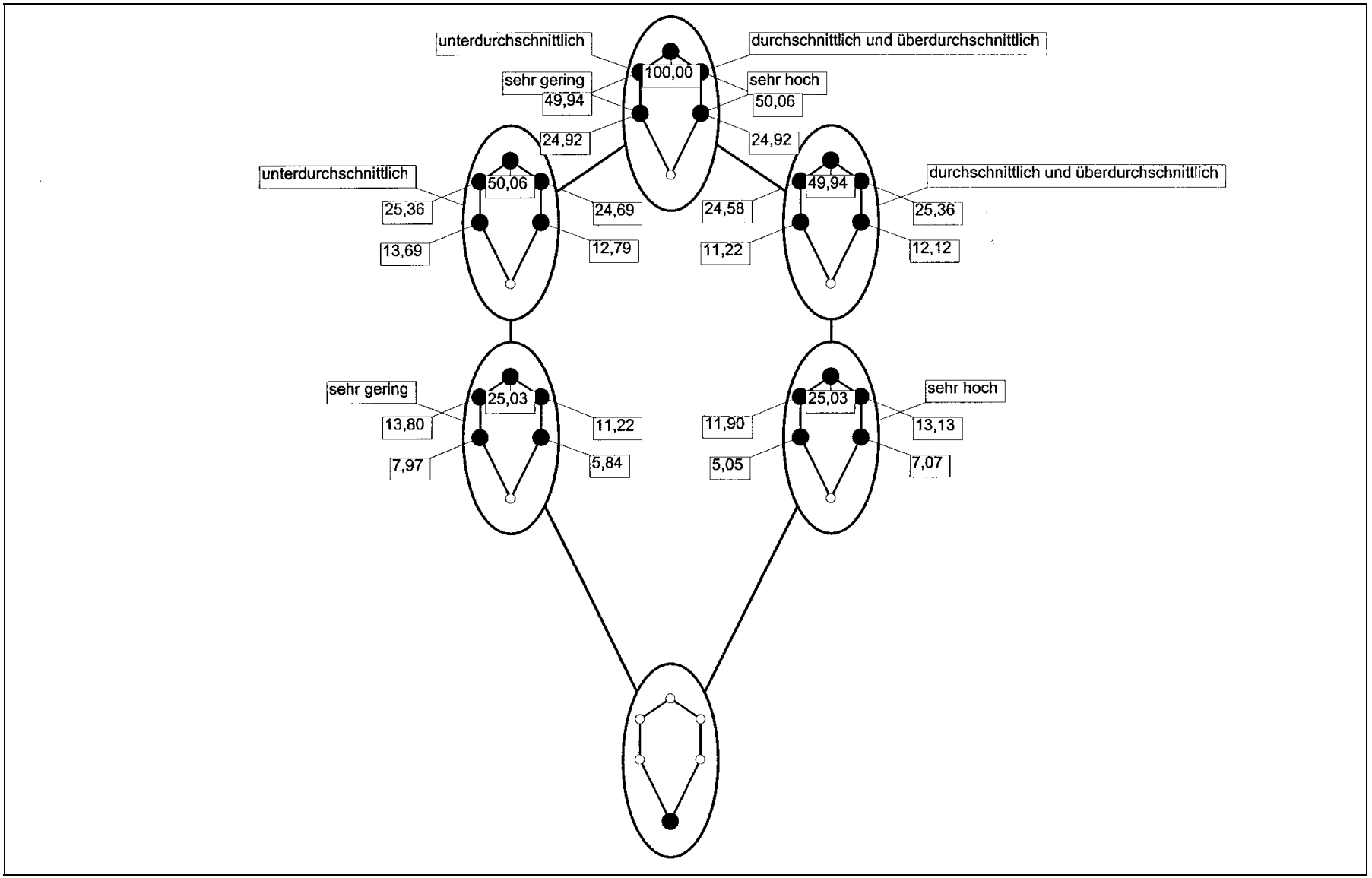


Abbildung B45: : Liniendiagramm für Reinertrag je AK und Lohn je entlohnte AK

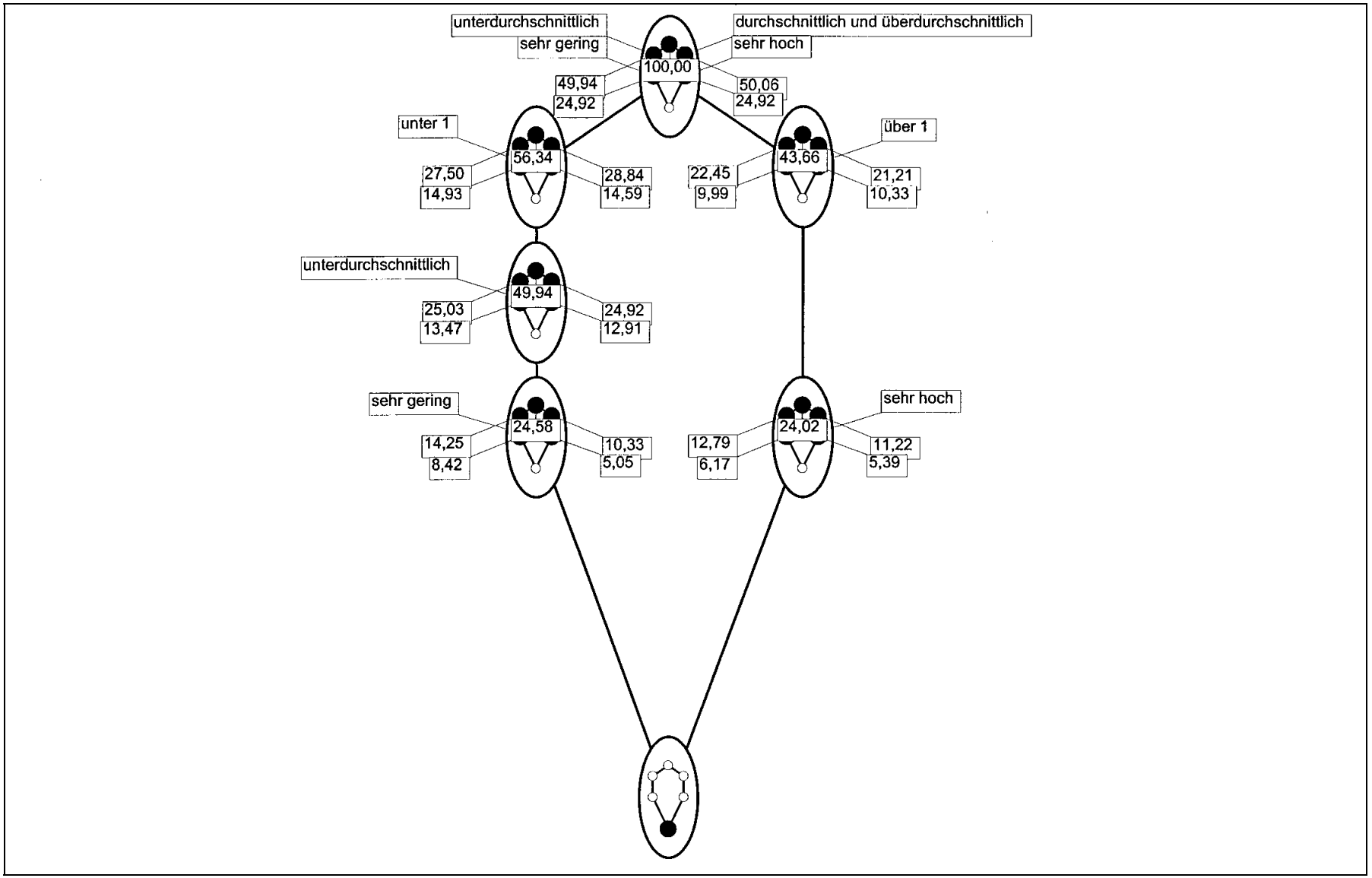


Abbildung B46: Liniendiagramm für Rentabilitätskoeffizient und Lohn je entlohnte AK

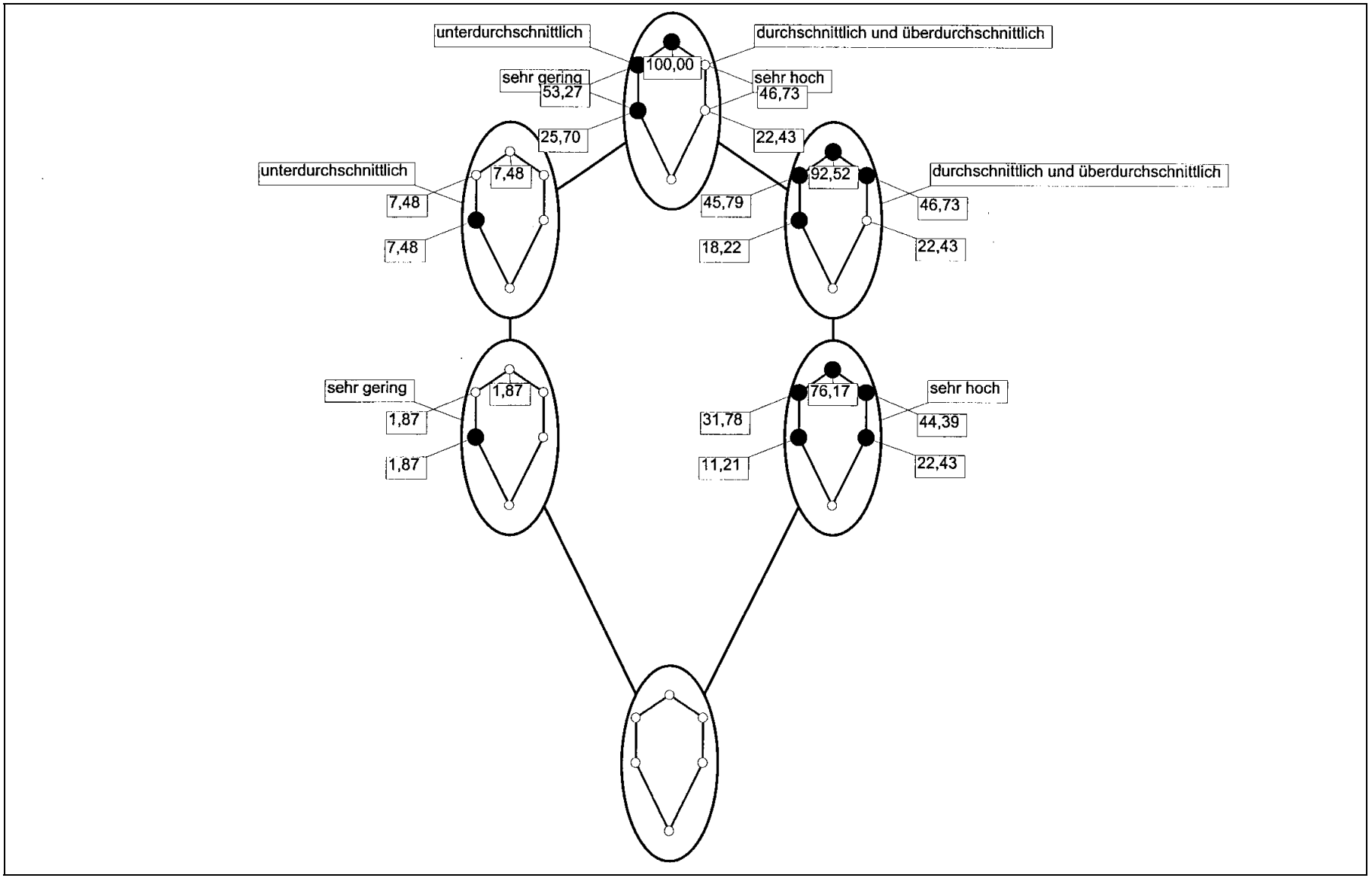


Abbildung B47: Liniendiagramm für Betriebseinkommen je AK und Lohn je entlohnte AK bei sehr hohem Rentabilitätskoeffizienten

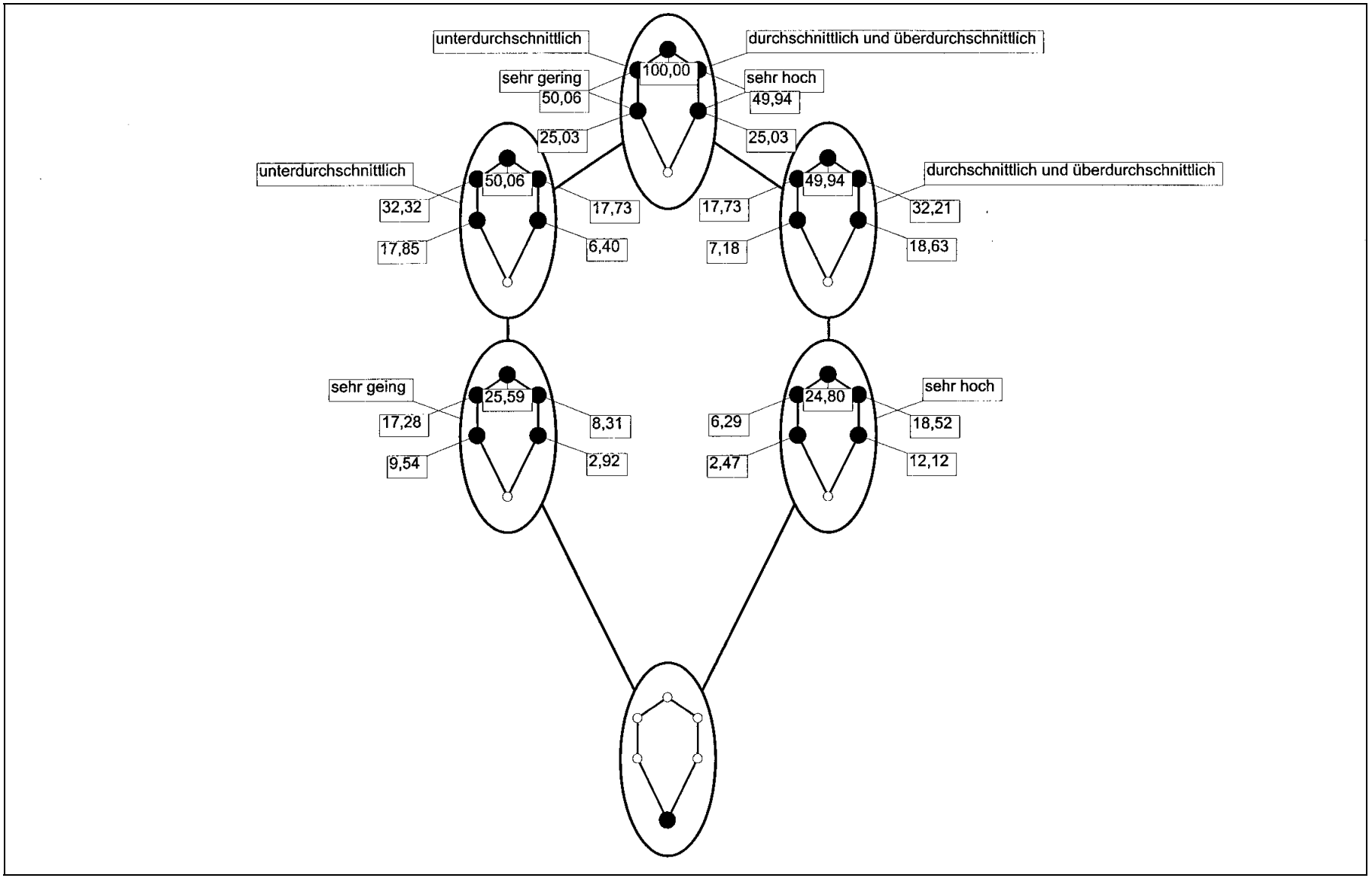


Abbildung B48: Liniendiagramm für Glasfläche je AK und Betriebseinkommen je AK

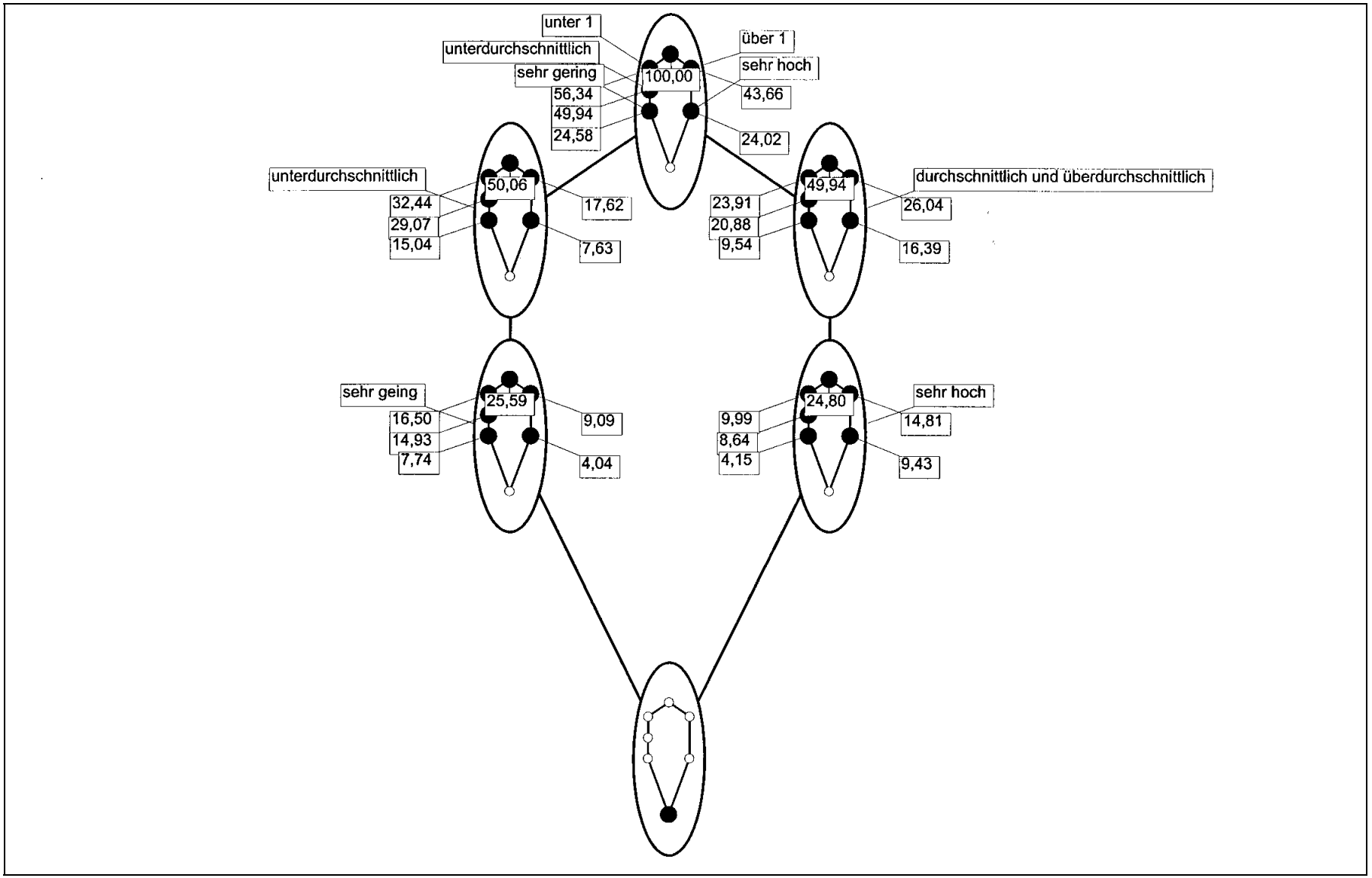


Abbildung B49: Liniendiagramm für Glasfläche je AK und Rentabilitätskoeffizient

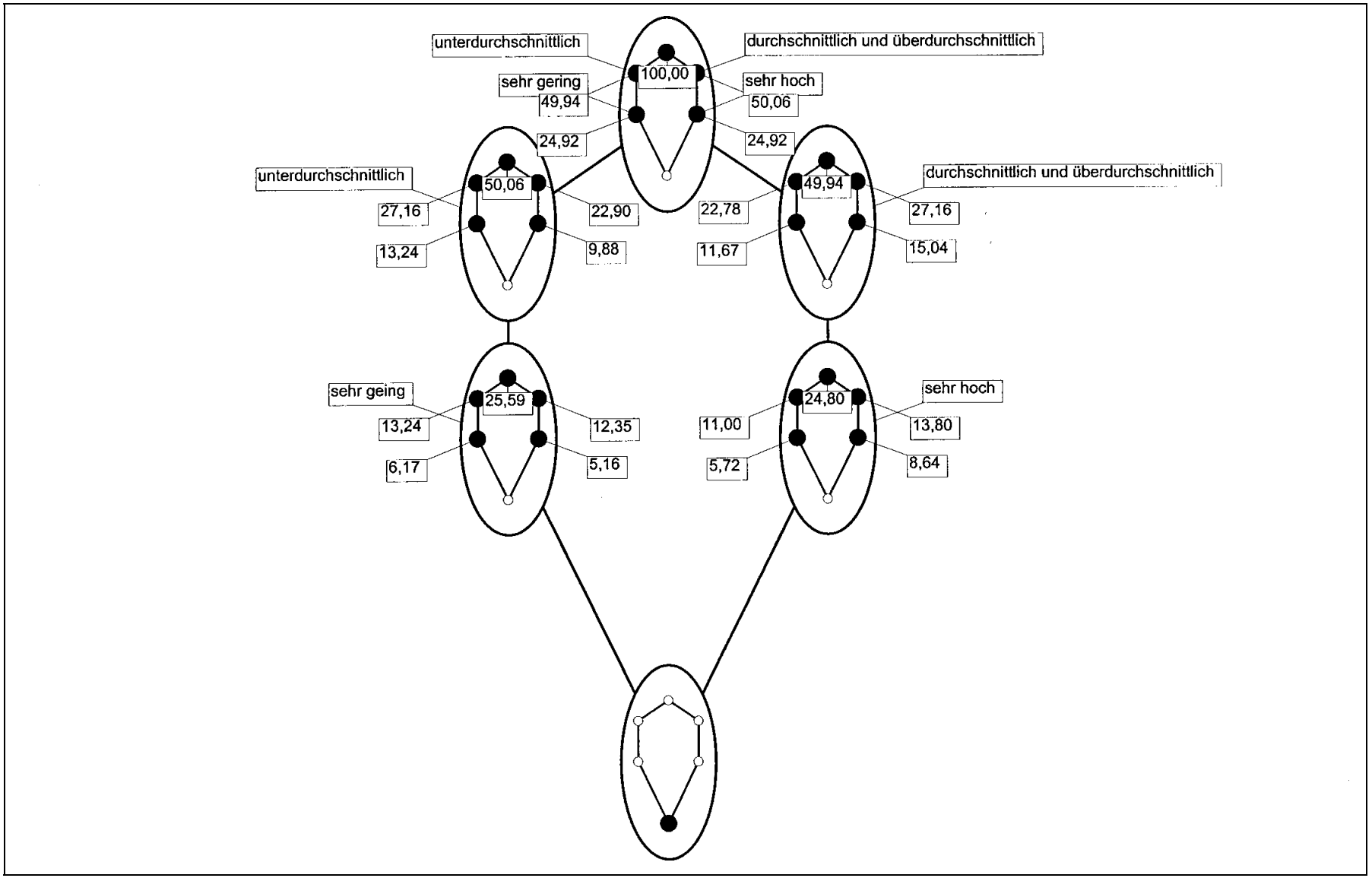


Abbildung B50: Liniendiagramm für Glasfläche je AK und Lohn je entlohnte AK

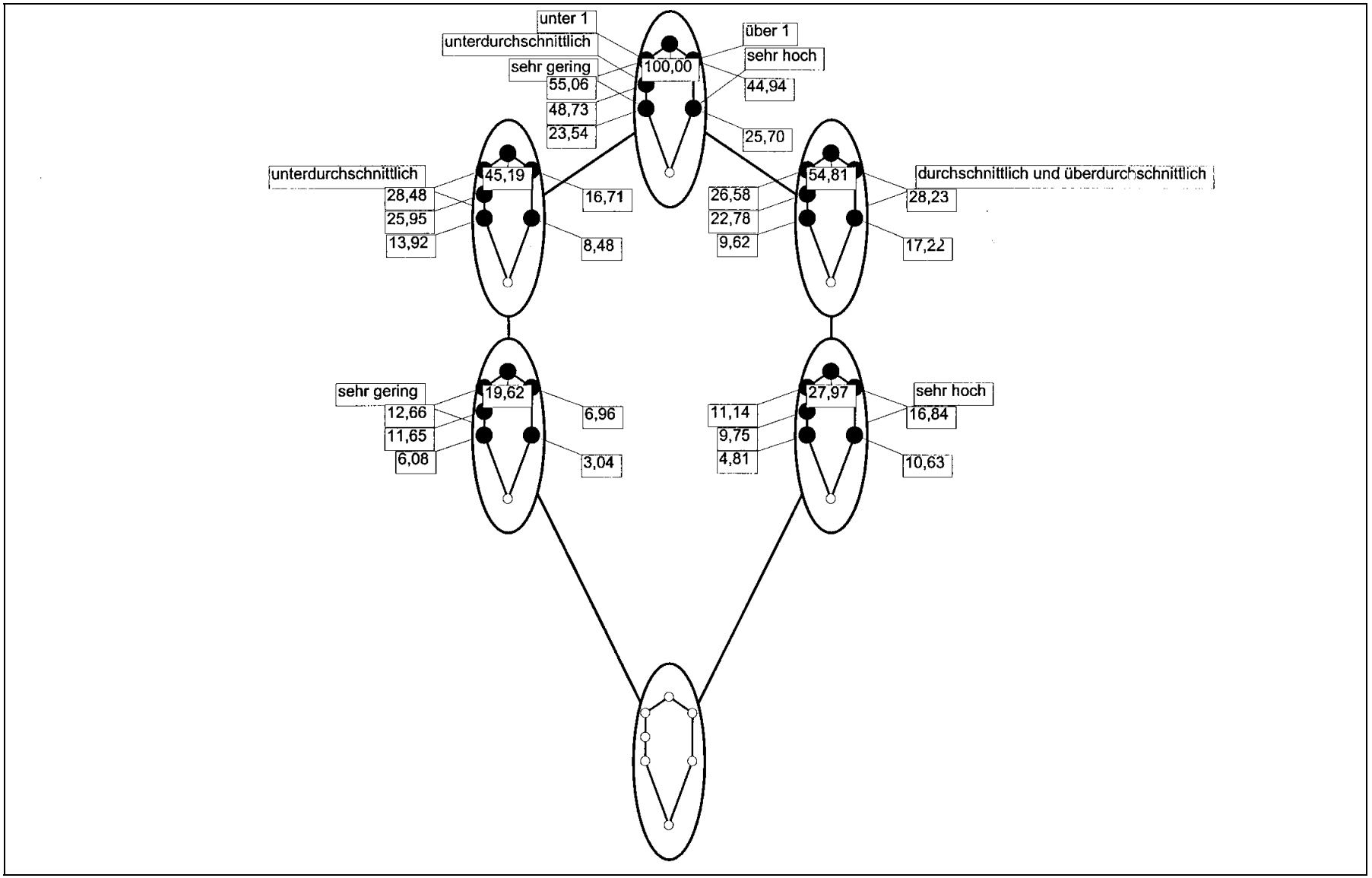


Abbildung B51: Liniendiagramm für Erträge aus Eigenproduktion und Rentabilitätskoeffizient, überwiegend indirekt absetzende Betriebe

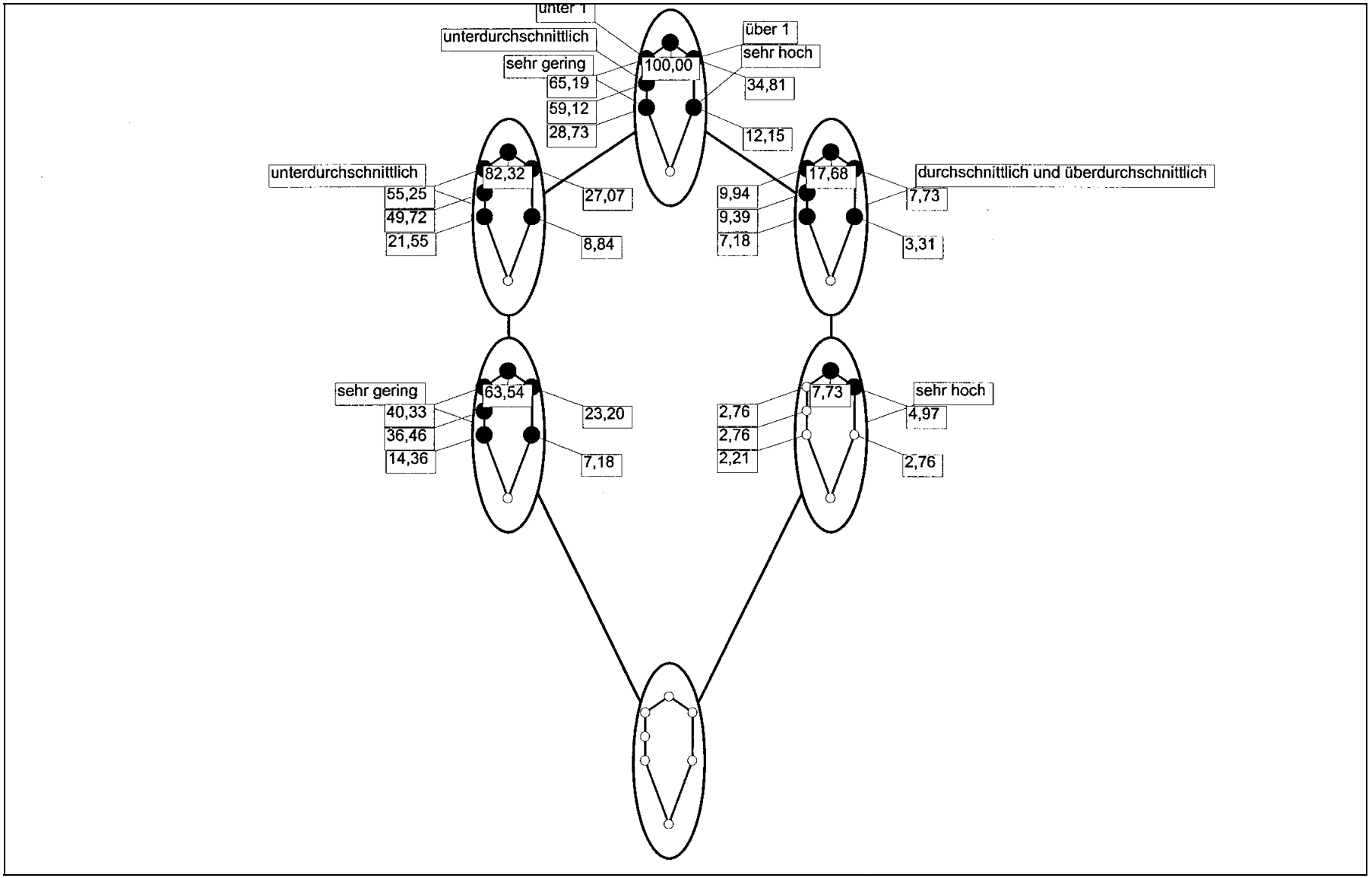


Abbildung B52: Liniendiagramm für Erträge aus Eigenproduktion und Rentabilitätskoeffizient, überwiegend direkt absetzende Betriebe

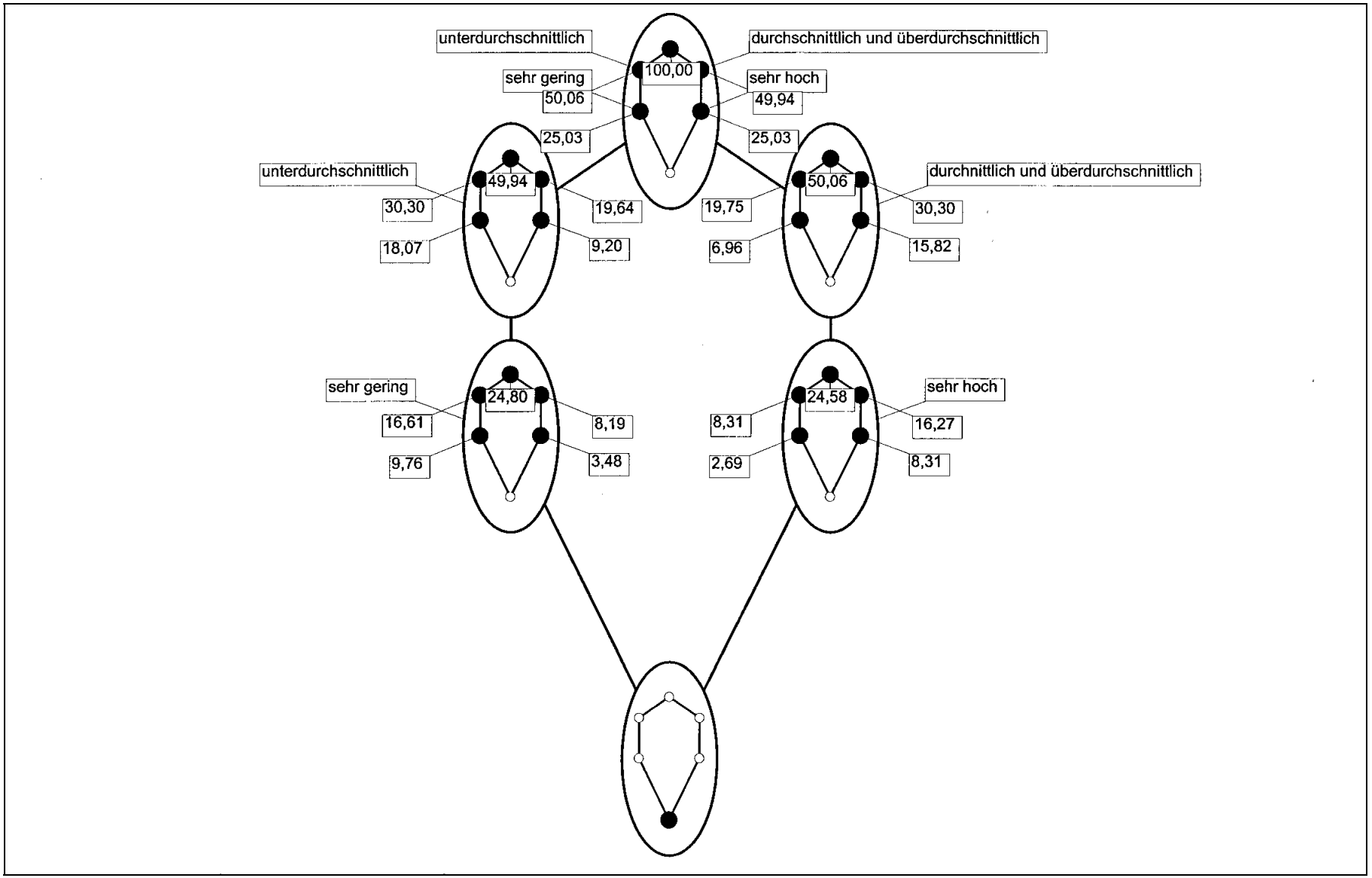


Abbildung B53: Liniendiagramm für Glasfläche in qm und Betriebseinkommen je AK

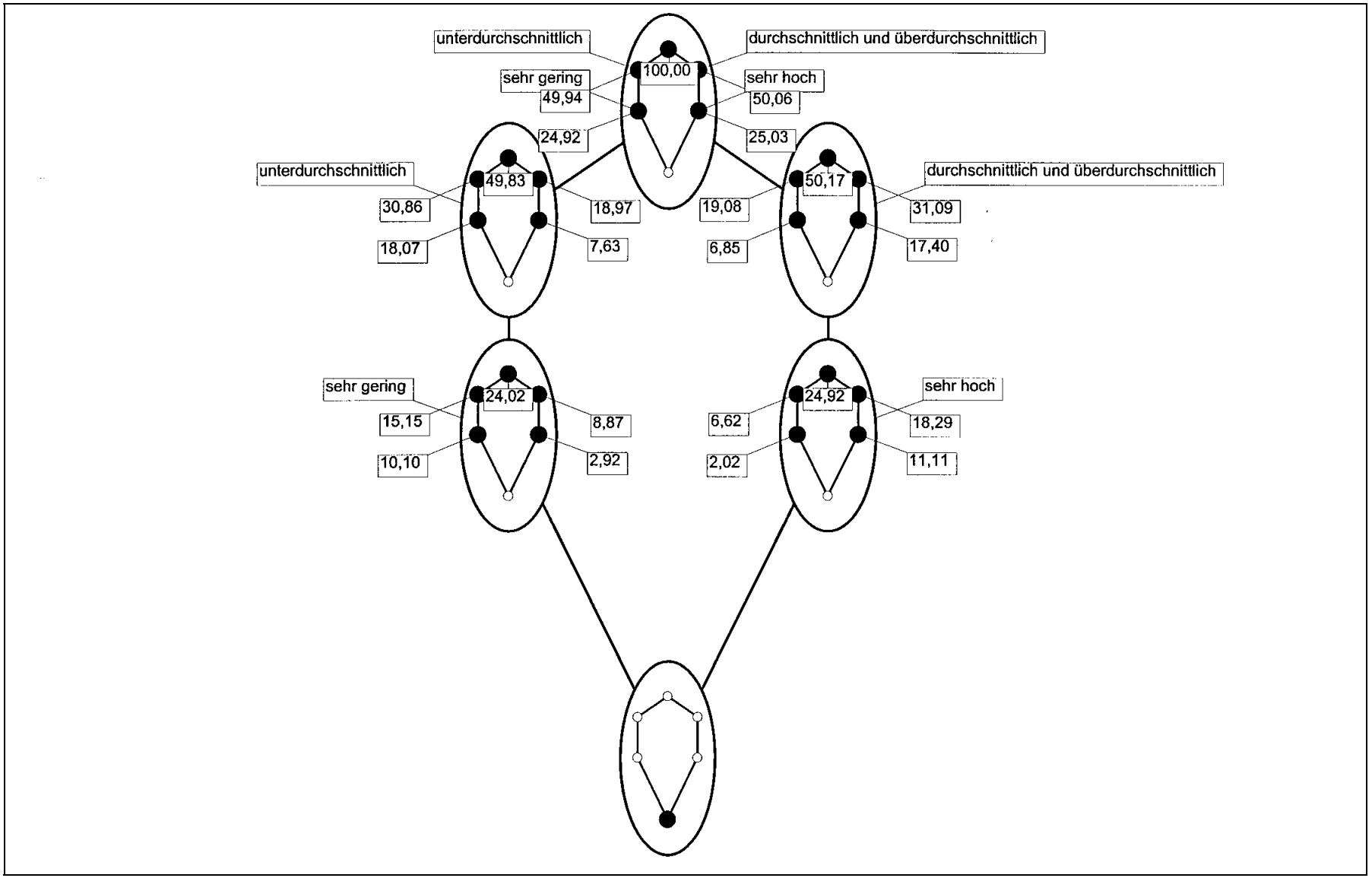


Abbildung B 54: Liniendiagramm für Arbeitskräfte insgesamt und Betriebseinkommen je Eqm

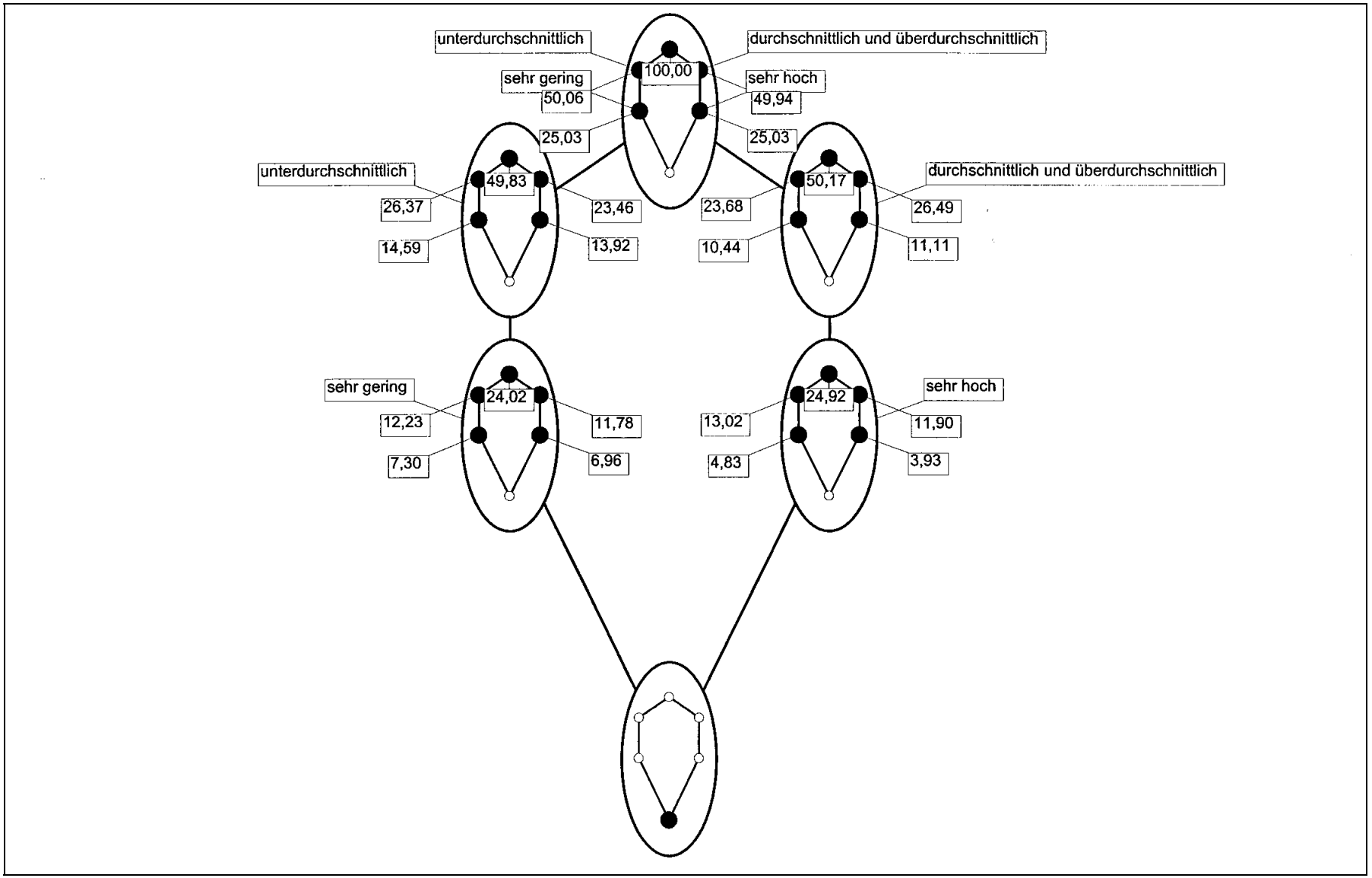


Abbildung B 55: Liniendiagramm für Arbeitskräfte insgesamt und Betriebseinkommen je AK

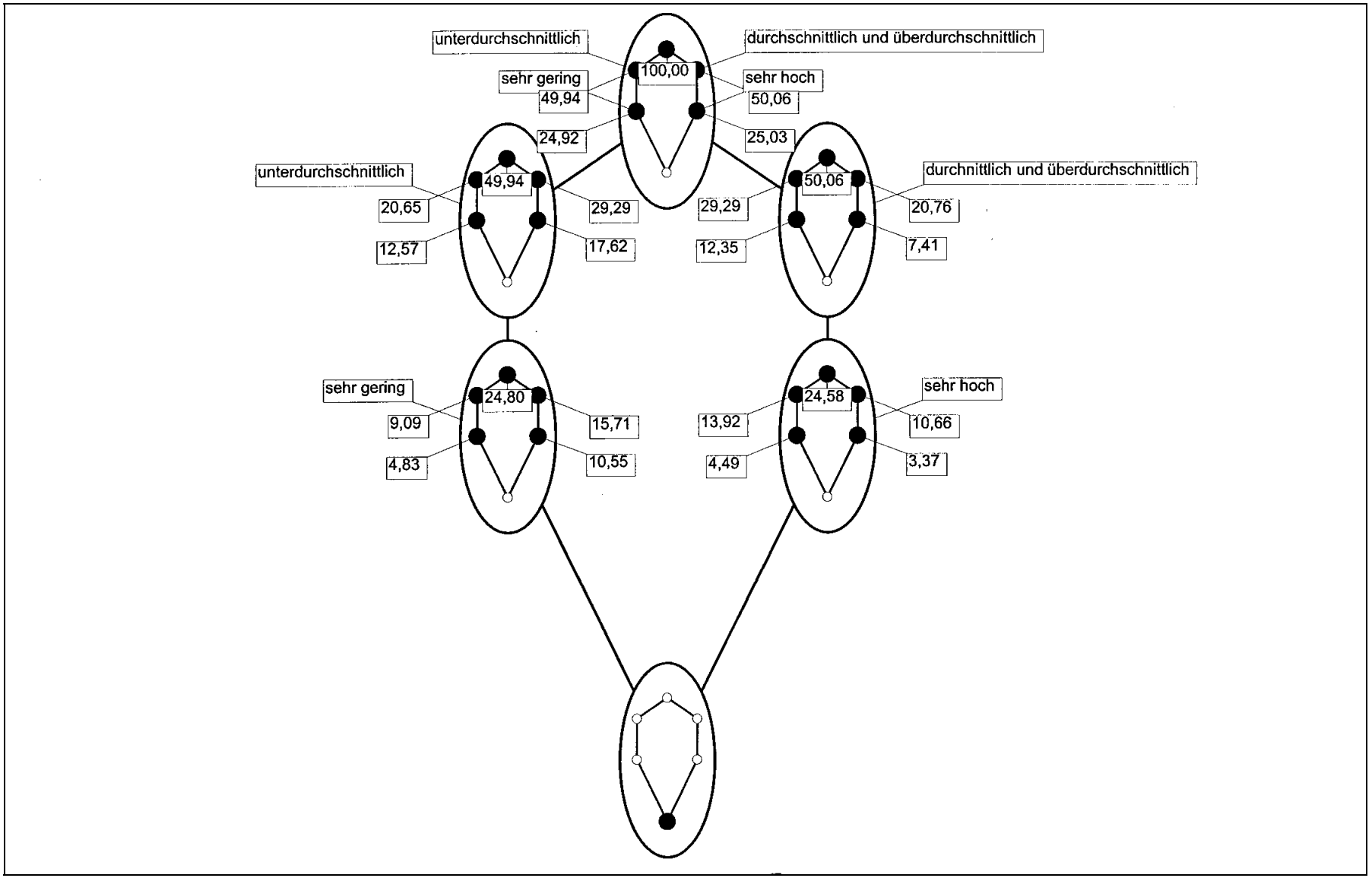


Abbildung B56: Liniendiagramm für Glasfläche in qm und Betriebseinkommen je Eqm

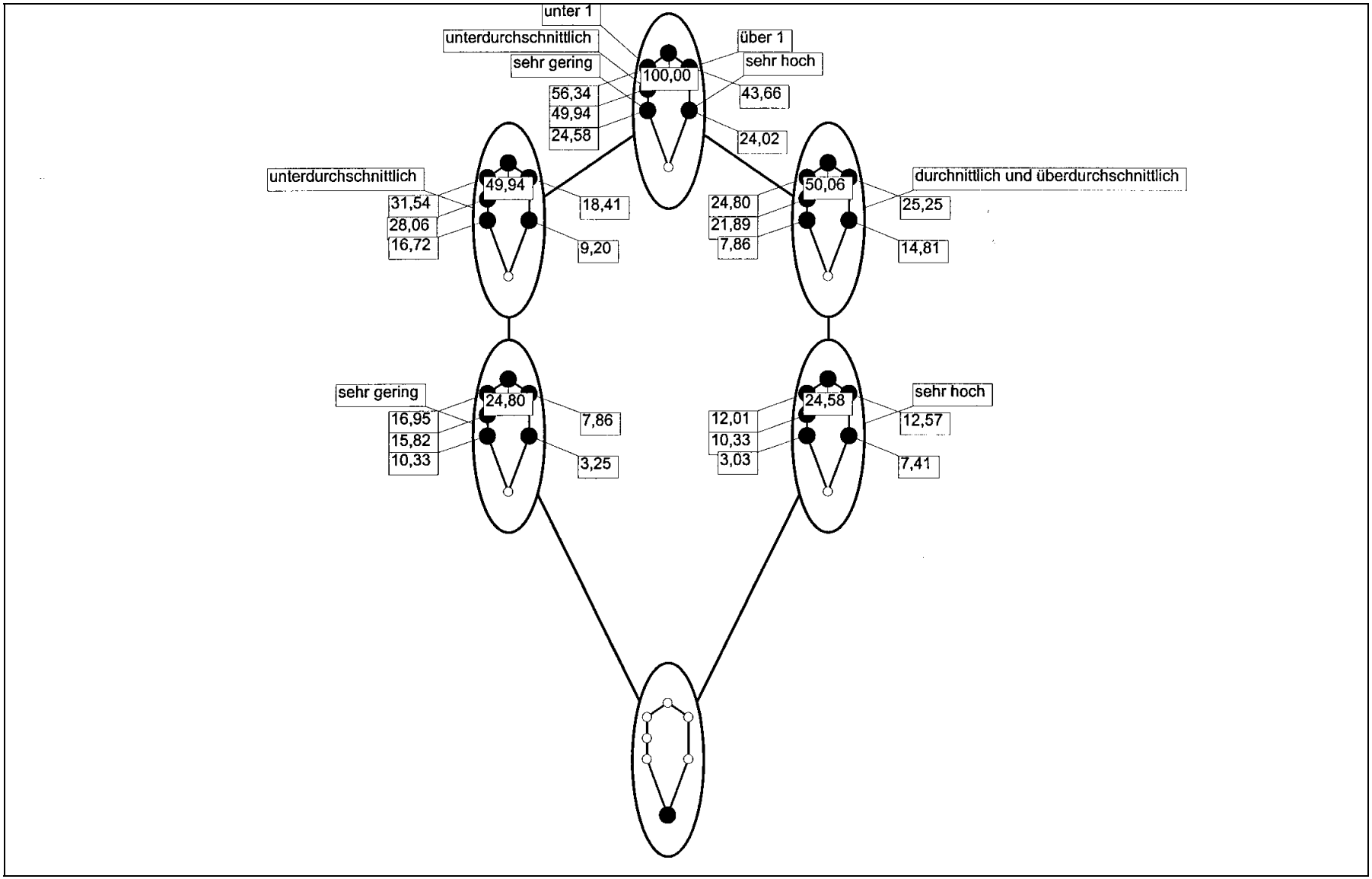


Abbildung B 57: Liniendiagramm für Glasfläche in qm und Rentabilitätskoeffizient

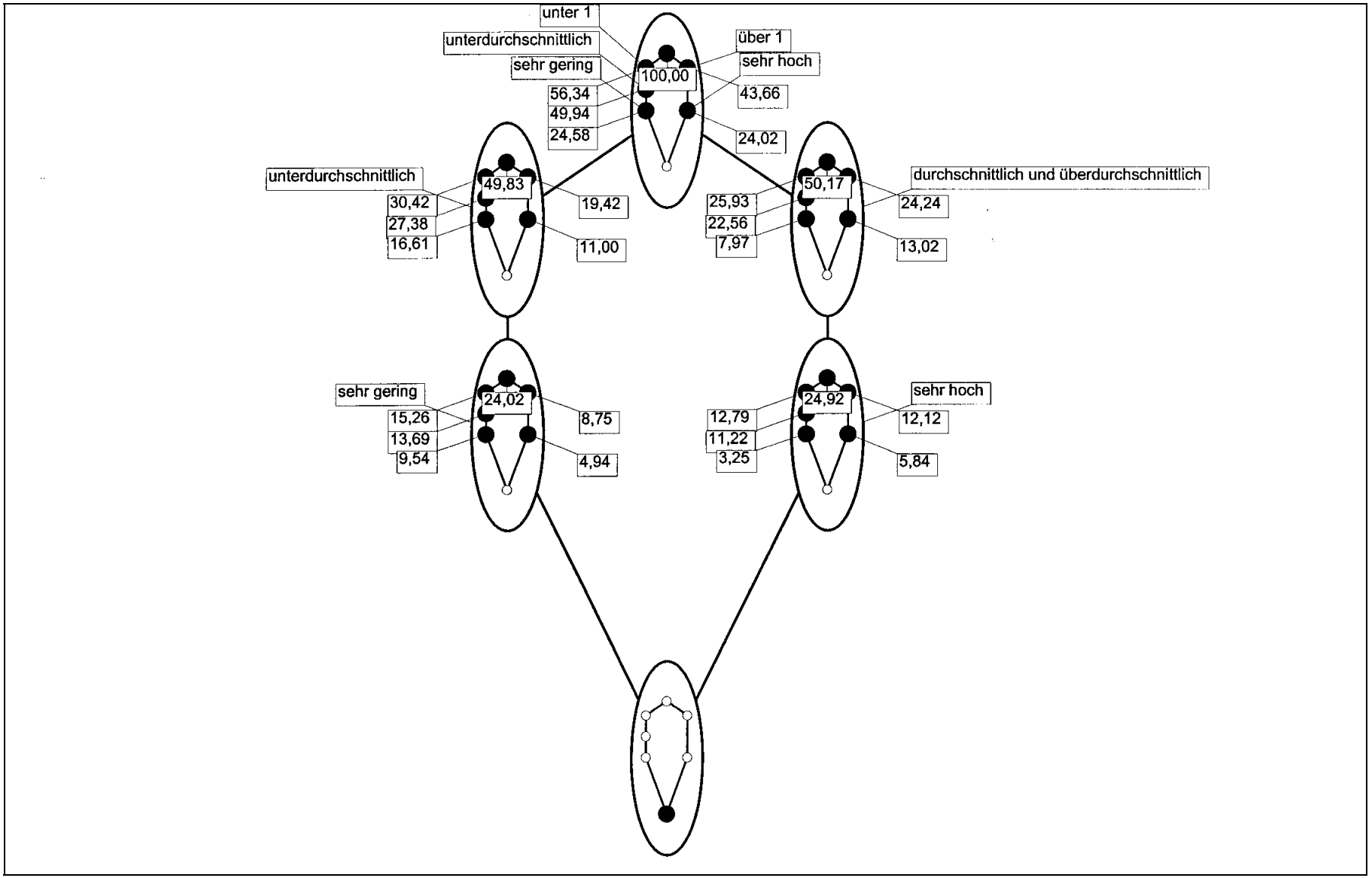


Abbildung B 58: Liniendiagramm für Arbeitskräfte insgesamt und Rentabilitätskoeffizient

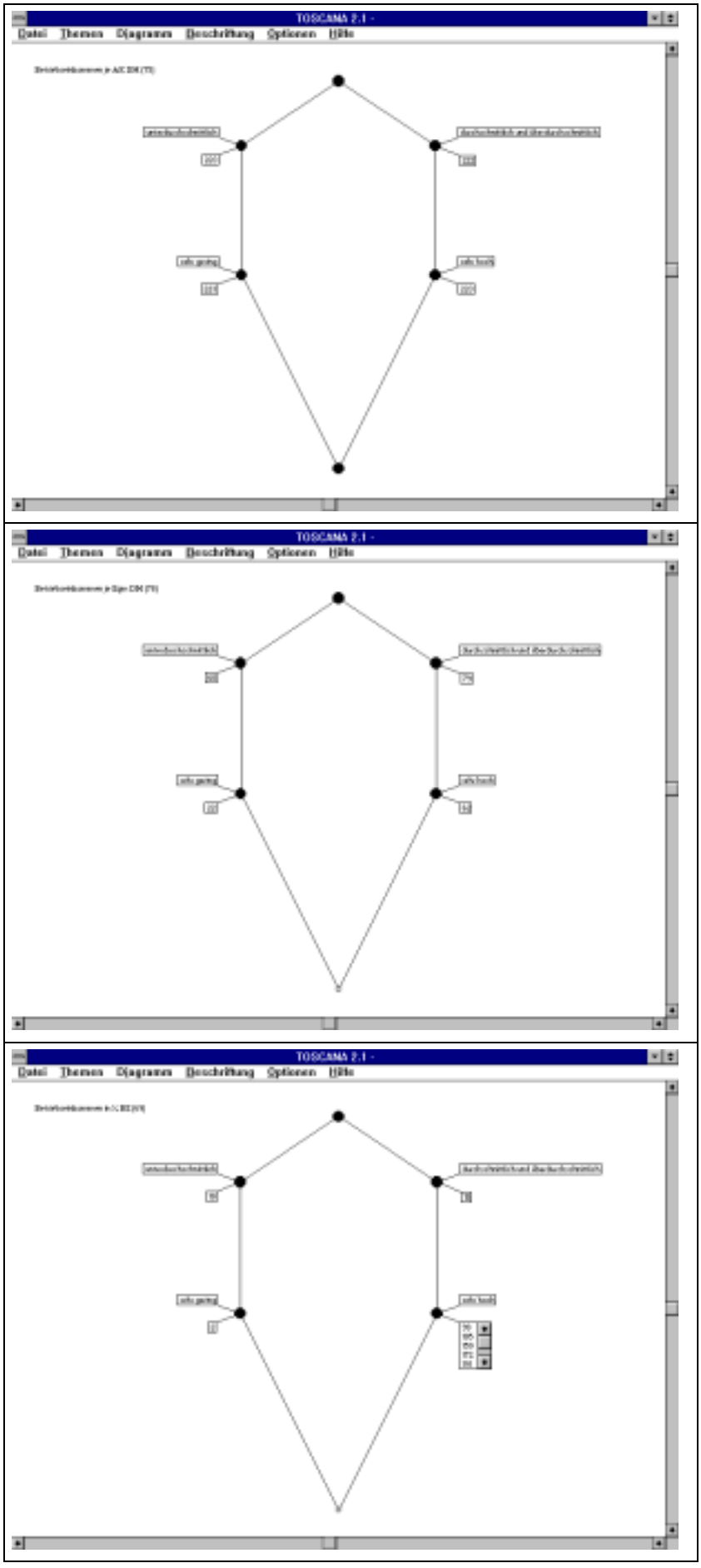


Abbildung B59: Der Weg durch Liniendiagramme zu der Gruppe von Betrieben mit sehr hoher Arbeits- und Flächenproduktivität und sehr hoher Wertschöpfungsquote

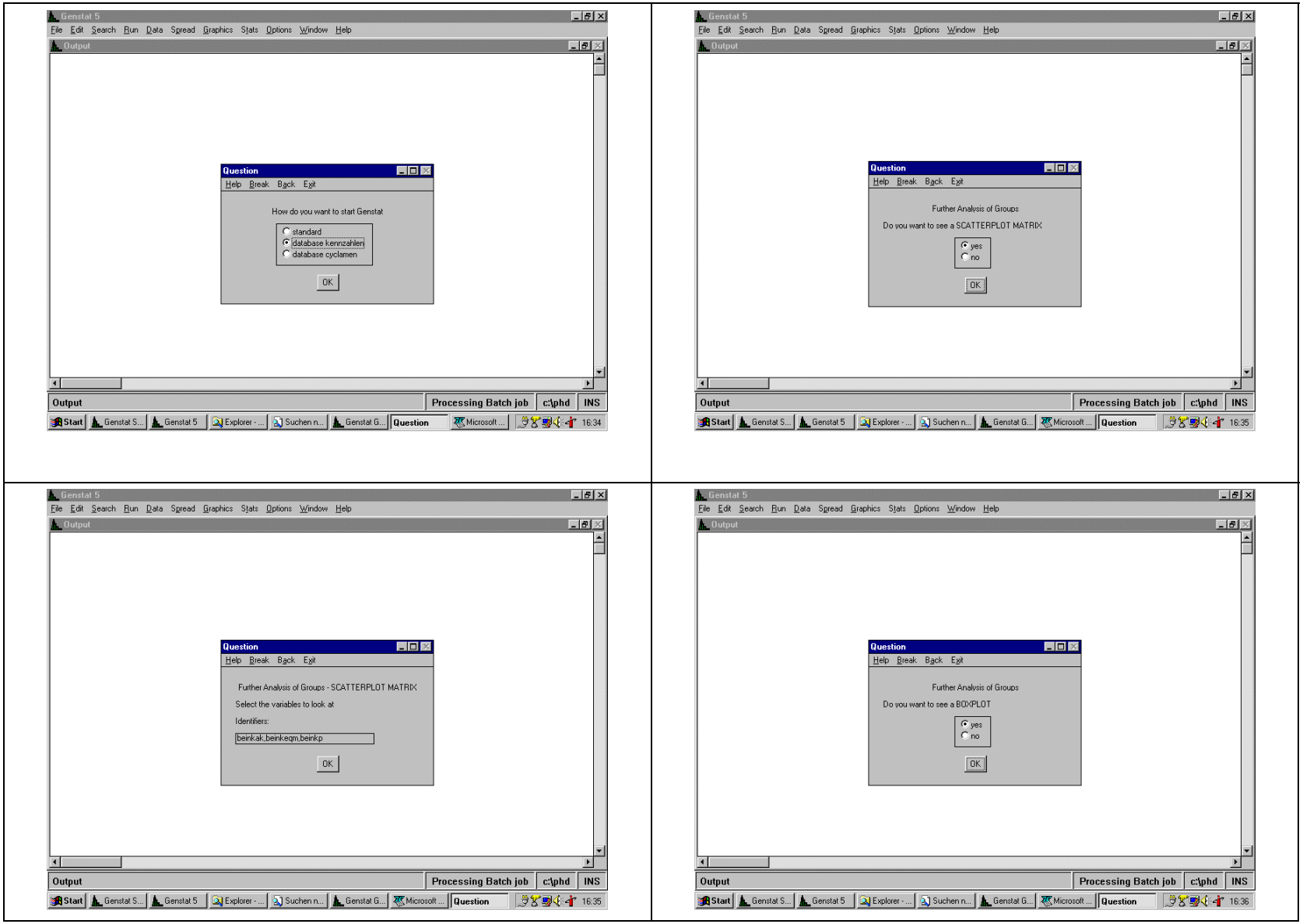


Abbildung B60a: Genstat Menüs zur Ergänzung der Analyse der Liniendiagramme

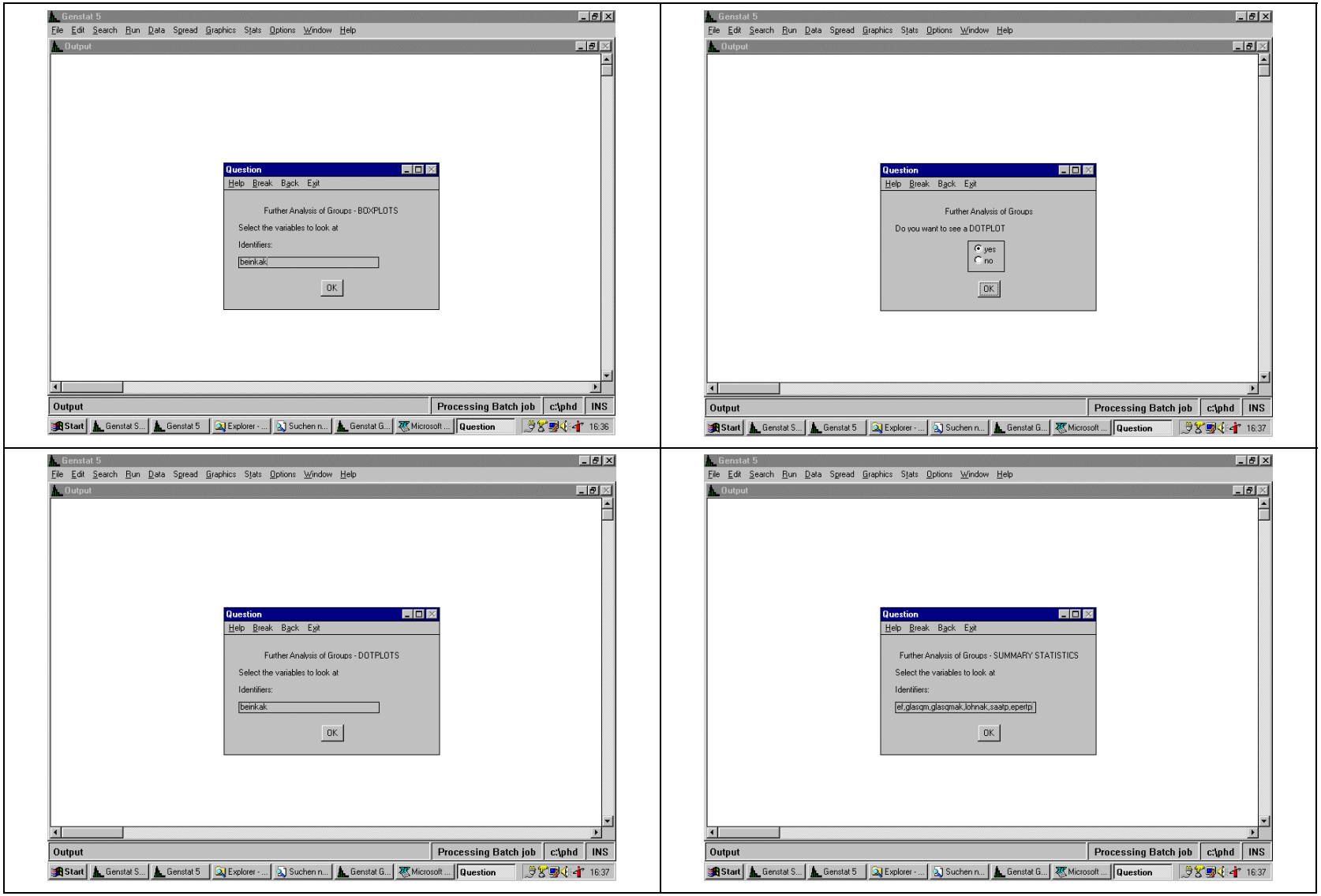
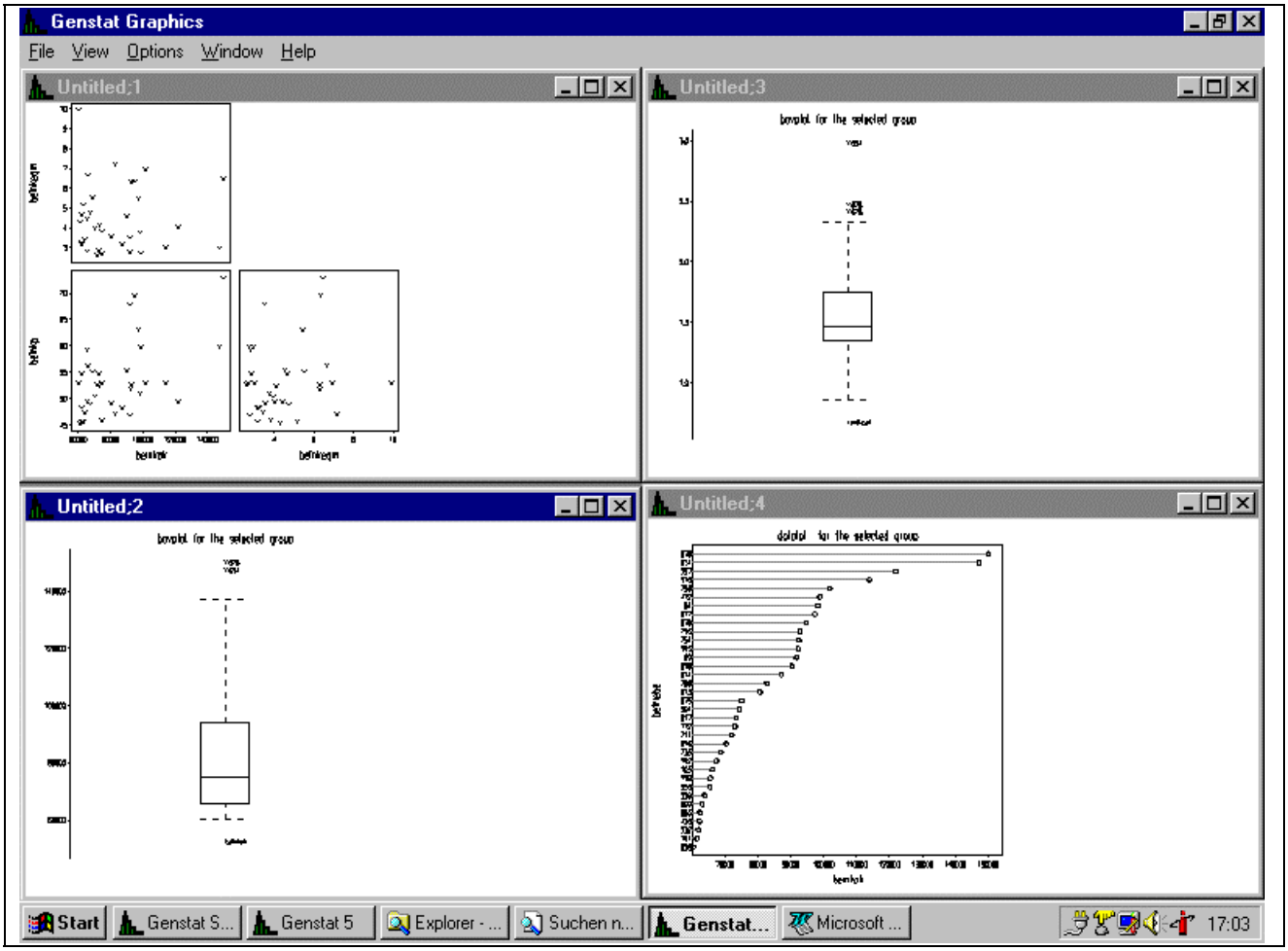


Abbildung B 60b: Genstat Menüs zur Ergänzung der Analyse der Liniendiagramme



summary statistics for
the whole data set

	length	mean	var.	median	min.	max.
rentkoef	891.0	0.9825	0.1624	0.9600	-3.120	2.980
glasqm	891.0	5769	12393471	4900	450.0	20000
glasqmak	891.0	1215	428444	1128	56.73	6656
lohnak	891.0	33134	160449781	32909	0	114511
saatp	891.0	11.75	117.7	9.770	0	159.3
epertp	891.0	92.72	78.80	96.84	61.54	100.0

summary statistics for
the selected group

	length	mean	var.	median	min.	max.
rentkoef	35.00	1.580	0.1822	1.460	0.8600	2.980
glasqm	35.00	4600	7609287	4500	843.0	12315
glasqmak	35.00	903.0	126256	767.1	303.2	2006
lohnak	35.00	40601	243509788	36723	16018	104593
saatp	35.00	9.117	39.00	8.580	0	22.15
epertp	35.00	92.02	109.6	97.41	62.69	99.97

Abbildung B61: Ergebnisausdruck der Genstat-Menüs aus Abbildung B60

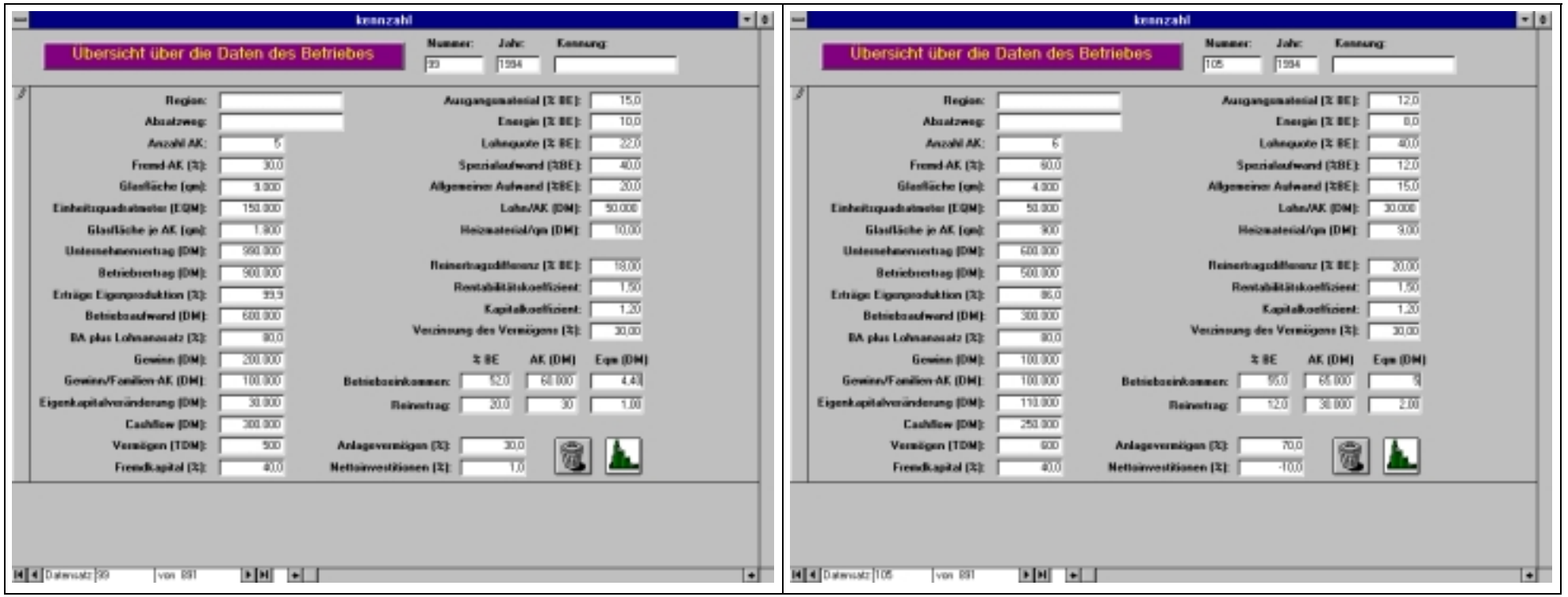


Abbildung B62: Zwei Betriebe des in Abbildung B 59 fokussierten Betriebes

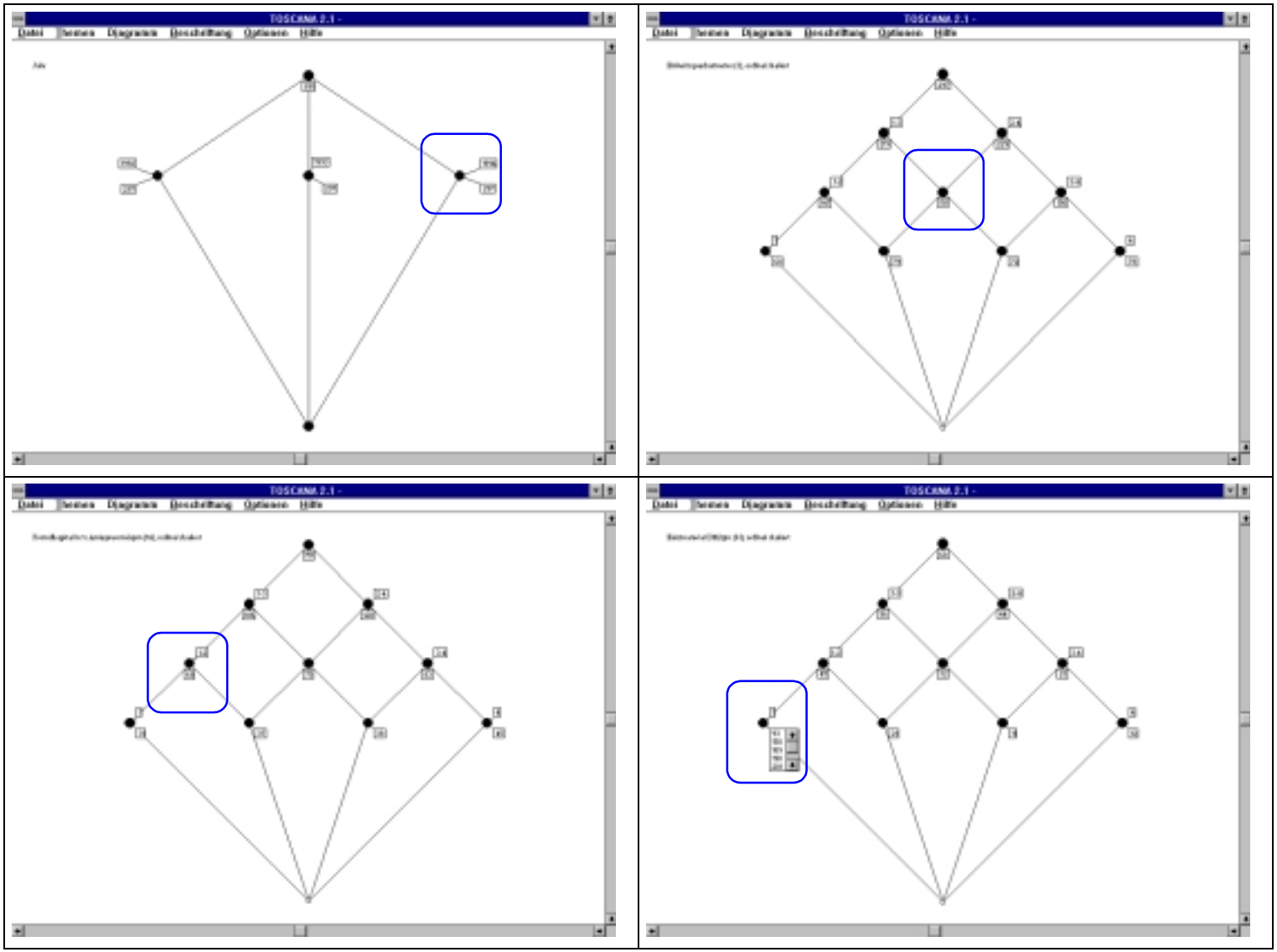


Abbildung B63: Der Weg durch ein Liniendiagramm zum Segment mit dem höchsten geschätzten Rentabilitätskoeffizienten 1994 in der CHAID-Analyse; jahr 1994, eqm Klasse 2 und 3, fkp Klasse 1 und 2, heizqm Klasse 1

Anhang Teil II A

Übersichten zur Auswertung der betriebsbegleitenden Untersuchung bei Cyclamen, Kapitel 3.1

Übersicht A1: Variablenset 1, Qualitätsbeurteilungen

Qualitätsbeurteilungen Sierra

	Anzahl	Gültiges N	Median	Spannweite
Gesamteindruck Woche 44	20	N=20	7,00	6,00
Gesamteindruck Woche 46	20	N=20	6,00	4,00
Gesamteindruck Woche 48	20	N=20	5,50	8,00
Vergilbung Woche 44	20	N=20	8,00	3,00
Vergilbung Woche 46	20	N=20	7,00	2,00
Vergilbung Woche 48	20	N=20	5,00	4,00
Knospenbesatz Woche 44	20	N=20	7,00	4,00
Knospenbesatz Woche 46	20	N=20	6,00	2,00
Knospenbesatz Woche 48	20	N=20	5,00	5,00
Krankheiten Woche 44	20	N=20	7,00	4,00
Krankheiten Woche 46	20	N=20	7,00	3,00
Krankheiten Woche 48	20	N=20	5,00	6,00
Welke Woche 44	20	N=20	7,00	4,00
Welke Woche 46	20	N=20	5,00	3,00
Welke Woche 48	20	N=20	5,00	4,00
Wurzelbild Woche 44	20	N=20	5,00	6,00
Wurzelbild Woche 48	20	N=20	5,00	4,00

Qualitätsbeurteilungen Concerto

	Anzahl	Gültiges N	Median	Spannweite
Gesamteindruck Woche 44	20	N=20	7,00	5,00
Gesamteindruck Woche 46	20	N=20	5,50	4,00
Gesamteindruck Woche 48	20	N=20	4,00	3,00
Vergilbung Woche 44	20	N=20	7,00	4,00
Vergilbung Woche 46	20	N=20	5,00	4,00
Vergilbung Woche 48	20	N=20	5,00	4,00
Knospenbesatz Woche 44	20	N=20	6,50	3,00
Knospenbesatz Woche 46	20	N=20	6,00	3,00
Knospenbesatz Woche 48	20	N=20	3,50	4,00
Krankeiten Woche 44	20	N=20	7,00	6,00
Krankheiten Woche 46	20	N=20	6,00	3,00
Krankheiten Woche 48	20	N=20	5,00	5,00
Welke Woche 44	20	N=20	6,00	6,00
Welke Woche 46	20	N=20	5,00	4,00
Welke Woche 48	20	N=20	4,00	4,00
Wurzelbild Woche 44	20	N=20	7,00	6,00
Wurzelbild Woche 48	20	N=20	5,00	8,00

Alle Merkmale sind auf einer Ordinalskala von 1 bis 9 bestimmt, wobei eine 1 immer für die schlechteste Beurteilung und eine 9 immer für die beste Beurteilung steht; das heißt, eine 9 im Knospenbesatz steht für eine Pflanze mit vielen Knospen und damit für gute Qualität. Eine 9 bei Krankheiten steht für eine Pflanze ohne Krankheiten oder Schädlinge, also ebenfalls für gute Qualität. Die Bonituren erfolgten eine Woche, vier Wochen und sechs Wochen nach dem Kulturende im Betrieb.
Für alle 20 Betriebe wurden 15 Pflanzen der Sorte ‘Sierra’ und 15 Pflanzen der Sorte ‘Concerto’ beurteilt. Verwendet wird für die Auswertungen der Median dieser Beurteilungen.
.

Übersicht A2: Variablenset 2, Substratanalysewerte

Substratanalysenwerte

	Anzahl	Gültiges N	Mittelwert	Standard-abweichung
K2O-Gehalt Woche 23, mg/l	20	N=20	239,95	75,02
K2O-Gehalt Woche 28, mg/l	20	N=20	183,65	94,08
K2O-Gehalt Woche 41, mg/l	20	N=19	129,32	117,67
N-Gehalt Woche 23, mg/l	20	N=20	174,25	66,02
N-Gehalt Woche 28, mg/l	20	N=20	153,30	74,21
N-Gehalt Woche 41, mg/l	20	N=19	115,84	205,60
pH-Wert Woche 23	20	N=20	5,91	,23
pH-Wert Woche 29	20	N=20	5,89	,34
pH-Wert Woche 41	20	N=19	5,88	,62
Salzgehalt Woche 23, g/l	20	N=20	1,18	,40
Salzgehalt Woche 29, g/l	20	N=20	1,38	,44
Salzgehalt Woche 41, g/l	20	N=19	1,77	,95

Bei den Substratanalysdewerten handelt es sich um verhältnisskalierte Variablen. Die Analysen wurden nach der LUFA-Methode in den Wochen 23, 28 und 41, das heißt zum Kulturbeginn, nach 6 Wochen Kulturdauer und zum Kulturende genommen. Für den Betrieb 5 fehlen die Substartanalysewerte in Woche 41.

Übersicht A 3a: Variablenset 3, Schattiersollwerte

Schattiersollwerte in den Wochen 23 bis 42 in Kilolux (klx)

	Anzahl	Gültiges N	Mittelwert	Standard-abweichung
Sollwert Woche 23	20	N=18	35,14	12,16
Sollwert Woche 24	20	N=18	35,97	11,56
Sollwert Woche 25	20	N=18	36,97	11,34
Sollwert Woche 26	20	N=18	37,69	10,99
Sollwert Woche 27	20	N=18	39,97	10,04
Sollwert Woche 28	20	N=18	39,97	10,04
Sollwert Woche 29	20	N=18	41,08	9,68
Sollwert Woche 30	20	N=18	41,36	9,72
Sollwert Woche 31	20	N=18	41,89	9,55
Sollwert Woche 32	20	N=18	41,89	9,55
Sollwert Woche 33	20	N=18	43,00	11,25
Sollwert Woche 34	20	N=18	43,97	13,17
Sollwert Woche 35	20	N=18	44,53	13,43
Sollwert Woche 36 - 42	20	N=18	44,53	13,43
Schattierfarbe ja-nein (nominalskaliert)	20	N=20		

Die Schattiersollwerte werden als verhältnisskaliert betrachtet. Bei der Angabe der Schattierfarbe handelt es sich um eine nominalskalierte Variable binärer Struktur.

Übersicht A 3b: Variablenset 3, Platzbedarf, Rücken

Kulturdaten - Platzbedarf, Rücken

	Anzahl	Gültiges N	Mittelwert	Median	Standardabweichung	Spannweite
Anzahl Rückvorgänge (nominalskaliert)	20	N=20				
aufgestellt mit ... Pflanzen/qm	20	N=20	58,75	64,00	15,88	58,00
Endstand (Pfl/qm)	20	N=20	14,35	14,00	3,34	15,00
Wochen auf Endstand	20	N=20	10,20	10,50	3,99	14,00
längste Phase mit einer Standweite (Wo)	20	N=20	5,65	5,00	1,50	6,00
Verhältnis Aufstellen:Enstand *	20	N=20	4,28	4,39	1,56	6,00
Verhältnis Stand vor und nach längster Phase *	20	N=20	2,32	2,00	,81	2,91
Nettowochenquadratmeter *	20	N=20	1040,31	1060,92	195,64	816,54

Die mit * gekennzeichneten Variablen sind aus anderen Variablen berechnet. Mit Ausnahme der Anzahl der Rückvorgänge, die als nominalskaliert betrachtet werden kann, und der berechneten Variablen, ist die Bestimmung des Skalenniveaus nicht ganz eindeutig und sowohl eine Interpretation der Variablen als verhältnis- als auch als ordinalskaliert denkbar.

Übersicht A 3c: Variablenset 3, Temperaturführung

Temperatureinstellungen

	Anzahl	Gültiges N	Mittelwert	Median	Standard-abweichung	Spannweite
negative Temperaturdifferenz Juni (nominalskaliert)	20	N=20				
negative Temperaturdifferenz Rest (nominalskaliert)	20	N=18				
Tagesmitteltemperatur Juni (Grad C)	20	N=20	16,93	17,00	1,23	5,00
Tagesmitteltemperatur Rest (Grad C)	20	N=18	15,78	16,00	1,24	4,25
Lüftungstemperatur Juni (Grad C)	20	N=20	19,41	19,50	1,43	5,00
Lüftungstemperatur Rest (Grad C)	20	N=20	17,88	18,00	1,62	7,00
Lüftung über Heizung Juni (Grad C) *	20	N=20	2,49	2,00	1,37	5,25
Lüftung über Heizung Rest (Grad C) *	20	N=18	2,25	2,00	1,17	4,25
geschätzter Energieverbrauch (l Heizöl/1000 Pflanzen) *	20	N=20	427,95	417,00	134,13	491,00

Bei den Temperaturwerten in Grad Celcius (C) handelt es sich um intervallskalierte Variablen. Bei den Angaben zur Temperaturdifferenz liegen nominalskalierte Variablen vor (keine negative Dif, ein Grad negative Dif, zwei Grad negative Dif), die auch ordinalskaliert interpretiert werden können. Die mit * gekennzeichneten Variablen sind aus den anderen Werten abgeleitet. Der Energiebverbrauch ist nach der im Planungsprogramm Gartplan verfügbaren Temperaturtabelle geschätzt (KRUSCHE, 1997). Der Energieverbrauch wird als verhältnisskaliert betrachtet.

Übersicht A4: Variablenset 4, Strukturdaten

ABS absatzwege Value Label 0 mehr als einer 1 einer BEW1 bewässerungsverfahren stellfläche 1 Value Label 0 über fuß (anstau, ebbe-flut, fließverfahren) 1 über kopf (gießwagen, von oben) BEW2 bewässerungsverfahren stellfläche 2 Value Label 0 über fuß (anstau, ebbe-flut, fließverfahren) 1 über kopf (gießwagen, von oben) GROE betriebsgröße (1 <, 2 >=) Value Label 0 kleiner 10000 qm 1 größer/gleich 10000 qm KREI region Value Label 0 östliches münsterland 1 westliches münsterland	MEN produktionsmenge Value Label 0 unter 50000 1 50000 und mehr SF1 stellfläche 1 Value Label 0 herkömmliche systeme (grundbeete, tische) 1 moderne tischsysteme (mobil, rinnen, roll) SF2 stellfläche 2 Value Label 0 herkömmliche systeme (grundbeete, tische) 1 moderne tischsysteme (mobill, rinnen, roll) SUBS substrat Value Label 1 andere 2 Einheitserden
Bei den Strukturvariablen handelt es sich um nominalskalierte Variablen binärer Struktur, das heißt es gibt für jedes Merkmal nur zwei Auprägungen. Die Ausgangsdaten führen die einzelnen Merkmale allerdings in mehr als zwei Klassen. Die Dichotomisierung wurde durch den Verfasser vorgenommen. Keine fehlenden Werte.	

Übersicht A5: Spearman Rangkorrelationen der Sorte ‘Sierra’ der Merkmale im Variablenset 1

```
Sample size:                20
Degrees of freedom =        18

Exact critical values for one-sided test:
 p=0.05, critical value = 0.377
 p=0.01, critical value = 0.534

*** Correlation matrix (adjusted for ties) ***

  1   1.000
  2   0.512   1.000
  3   0.282   0.470   1.000
  4   0.341   0.213   0.325   1.000
  5   0.345   0.056   0.055   0.446   1.000
  6  -0.010   0.075   0.389   0.478   0.174   1.000
  7   0.087   0.014  -0.009  -0.239  -0.113   0.064   1.000
  8   0.302   0.480   0.348   0.000   0.000  -0.226  -0.053   1.000
  9   0.389   0.071  -0.011   0.050   0.091  -0.075   0.176   0.280   1.000
 10   0.136  -0.053  -0.036   0.047   0.094   0.030  -0.400  -0.015   0.343   1.000
 11   0.456   0.327   0.387   0.221   0.327   0.309  -0.268   0.000   0.263   0.687   1.000
 12  -0.041   0.371   0.551   0.392   0.331   0.533  -0.139   0.053  -0.150   0.106   0.448   1.000
 13   0.325  -0.030   0.200   0.318   0.254   0.293  -0.088  -0.232   0.122   0.546   0.469   0.349   1.000
 14   0.137   0.016   0.232   0.200   0.399   0.429  -0.002  -0.257  -0.138   0.139   0.356   0.571   0.599   1.000
 15  -0.105   0.070   0.051  -0.050   0.048   0.270   0.153  -0.117  -0.031   0.151   0.258   0.382   0.318   0.498   1.000
 16   0.295   0.191   0.109  -0.003   0.033  -0.058   0.351   0.228  -0.024   0.509  -0.176  -0.311  -0.381  -0.048   0.105   1.000
 17   0.141   0.217   0.348  -0.211  -0.038   0.118  -0.009   0.197   0.238   0.093   0.259   0.102   0.120  -0.061   0.225  -0.079  1.000

          1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17

   Reihenfolge der Variablen:

   1 ges 44     7 kno 44   13 wel 44
   2 ges 46     8 kno 46   14 wel 46
   3 ges 48     9 kno 48   15 wel 48
   4 gil 44    10 kra 44   16 wur 44
   5 gil 46    11 kra 46   17 wur 48
   6 gil 48    12 kra 48
```

Übersicht A6: Spearman Rangkorrelationen der Sorte ‘Concerto’ der Merkmale im Variablenset 1

Spearman Rank Correlation

Sample size: 20
Degrees of freedom = 18

Exact critical values for one-sided test:
p=0.05, critical value = 0.377
p=0.01, critical value = 0.534

*** Correlation matrix (adjusted for ties) ***

1	1.000															
2	0.336	1.000														
3	0.673	0.482	1.000													
4	0.592	0.370	0.292	1.000												
5	0.081	0.450	0.260	0.486	1.000											
6	0.209	0.313	0.410	0.564	0.739	1.000										
7	0.695	0.309	0.452	0.286	0.080	0.082	1.000									
8	-0.048	0.434	0.297	0.189	0.786	0.655	0.084	1.000								
9	-0.207	0.198	-0.473	-0.086	0.181	-0.229	-0.251	-0.007	1.000							
10	0.611	0.260	0.354	0.603	0.131	0.331	0.442	0.212	-0.306	1.000						
11	0.556	0.406	0.382	0.580	0.605	0.529	0.265	0.540	0.067	0.659	1.000					
12	0.527	-0.125	0.170	0.288	-0.070	0.138	0.476	-0.218	-0.278	0.456	0.277	1.000				
13	0.764	0.555	0.548	0.521	0.242	0.270	0.828	0.242	-0.153	0.627	0.519	0.359	1.000			
14	0.072	0.285	0.202	-0.275	0.096	-0.050	0.205	0.223	0.098	-0.092	0.024	-0.034	0.231	1.000		
15	0.220	0.337	0.153	0.015	0.316	0.199	0.365	0.443	0.026	0.239	0.456	0.116	0.384	0.757	1.000	
16	-0.218	-0.225	-0.103	-0.018	-0.232	-0.091	-0.127	-0.124	-0.383	-0.097	-0.221	0.000	-0.164	-0.573	-0.378	1.000
17	-0.137	-0.029	-0.073	0.066	0.265	0.004	-0.191	0.163	0.139	-0.351	-0.032	-0.096	-0.096	-0.123	-0.239	0.268	1.000

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Reihenfolge der Variablen:

1 ges 44	7 kno 44	13 wel 44
2 ges 46	8 kno 46	14 wel 46
3 ges 48	9 kno 48	15 wel 48
4 gil 44	10 kra 44	16 wur 44
5 gil 46	11 kra 46	17 wur 48
6 gil 48	12 kra 48	

Übersicht A7: Eigenwerte und Spur der Hauptkoordinatenanalysen der Qualitätsbonituren für ‘Sierra’ und ‘Concerto’ Woche 44 und Woche 48

‘Sierra’ 44

```
 *****  Principal coordinates analysis  *****
  ***  Latent Roots  ***
             1          2          3          4          5          6
        1.0246     0.7690     0.5554     0.4250     0.3054     0.2236
             7          8          9         10         11         12
        0.2087     0.1608     0.1326     0.1161     0.0801     0.0474
            13         14         15         16         17         18
        0.0392     0.0273     0.0191     0.0162     0.0101     0.0061
            19         20
        0.0000     0.0000
  ***  Percentage variation  ***
             1          2          3          4          5          6
         24.59      18.46      13.33      10.20       7.33       5.37
             7          8          9         10         11         12
         5.01       3.86       3.18       2.79       1.92       1.14
            13         14         15         16         17         18
         0.94       0.66       0.46       0.39       0.24       0.15
            19         20
         0.00       0.00
***  Trace  4.167
```

‘Sierra’ 48

```
 *****  Principal coordinates analysis  *****
  ***  Latent Roots  ***
             1          2          3          4          5          6
        0.8335     0.5675     0.4028     0.3091     0.2980     0.2043
             7          8          9         10         11         12
        0.1656     0.1175     0.1018     0.0851     0.0658     0.0400
            13         14         15         16         17         18
        0.0330     0.0254     0.0079     0.0072     0.0000     0.0000
            19         20
        0.0000     0.0000
***  Percentage variation  ***
             1          2          3          4          5          6
         25.53      17.38      12.34       9.47       9.13       6.26
             7          8          9         10         11         12
         5.07       3.60       3.12       2.61       2.02       1.23
            13         14         15         16         17         18
         1.01       0.78       0.24       0.22       0.00       0.00
            19         20
         0.00       0.00
***  Trace  3.265
```

‘Concerto’ 44

```
 *****  Principal coordinates analysis  *****
 ***  Latent Roots  ***
             1          2          3          4          5          6
        1.5171     0.6916     0.4721     0.3354     0.3258     0.2353
             7          8          9         10         11         12
        0.2107     0.1698     0.1343     0.0680     0.0579     0.0426
            13         14         15         16         17         18
        0.0391     0.0237     0.0181     0.0174     0.0077     0.0045
            19         20
        0.0015     0.0000
 ***  Percentage variation  ***
             1          2          3          4          5          6
         34.70      15.82      10.80       7.67       7.45       5.38
             7          8          9         10         11         12
         4.82       3.88       3.07       1.55       1.32       0.97
            13         14         15         16         17         18
         0.89       0.54       0.41       0.40       0.18       0.10
            19         20
         0.03       0.00
***  Trace  4.372
```

‘Concerto’ 48

```
 *****  Principal coordinates analysis  *****
 ***  Latent Roots  ***
             1          2          3          4          5          6
        0.8644     0.4891     0.4197     0.2973     0.2551     0.2163
             7          8          9         10         11         12
        0.1821     0.1462     0.1097     0.0819     0.0761     0.0595
            13         14         15         16         17         18
        0.0491     0.0334     0.0290     0.0147     0.0093     0.0040
            19         20
        0.0027     0.0000
  ***  Percentage variation  ***
             1          2          3          4          5          6
         25.88      14.64      12.57       8.90       7.64       6.48
             7          8          9         10         11         12
         5.45       4.38       3.29       2.45       2.28       1.78
            13         14         15         16         17         18
         1.47       1.00       0.87       0.44       0.28       0.12
            19         20
         0.08       0.00
***  Trace  3.340
```

Übersicht A8: Spearman Korrelationsmatrizen für Bonituren der Qualitätsmerkmale für ‘Sierra’ und ‘Concerto’ in Woche 44 und 48

‘Sierra’, Woche 44

*** Correlation matrix (adjusted for ties) ***

1	1.000					
2	0.087	1.000				
3	0.295	0.351	1.000			
4	0.341	-0.239	-0.003	1.000		
5	0.325	-0.088	-0.381	0.318	1.000	
6	0.136	-0.400	-0.509	0.047	0.546	1.000
	1	2	3	4	5	6

‘Sierra’, Woche 48

*** Correlation matrix (adjusted for ties) ***

1	1.000					
2	-0.011	1.000				
3	0.348	0.238	1.000			
4	0.389	-0.075	0.118	1.000		
5	0.051	-0.031	0.225	0.270	1.000	
6	0.551	-0.150	0.102	0.533	0.382	1.000
	1	2	3	4	5	6

‘Concerto’, Woche 44

*** Correlation matrix (adjusted for ties) ***

1	1.000					
2	0.695	1.000				
3	-0.218	-0.127	1.000			
4	0.592	0.286	-0.018	1.000		
5	0.764	0.828	-0.164	0.521	1.000	
6	0.611	0.442	-0.097	0.603	0.627	1.000
	1	2	3	4	5	6

‘Concerto’, Woche 48

*** Correlation matrix (adjusted for ties) ***

1	1.000					
2	-0.473	1.000				
3	-0.073	0.139	1.000			
4	0.410	-0.229	0.004	1.000		
5	0.153	0.026	-0.239	0.199	1.000	
6	0.170	-0.278	-0.096	0.138	0.116	1.000
	1	2	3	4	5	6

‘Sierra’-’Concerto’ Kombinationen, Woche 44

Si 1	**0.474**	0.335	0.133	0.051	0.104	-0.225
Si 2	0.304	**0.054**	0.221	-0.134	-0.124	-0.059
Si 3	0.170	-0.003	**0.192**	0.069	0.114	0.071
Si 4	0.351	0.082	-0.351	**0.062**	0.315	0.094
Si 5	0.277	0.227	0.002	-0.177	**0.118**	0.082
Si 6	0.320	0.164	-0.022	0.019	0.056	**0.019**
	Co 1	Co 2	Co 3	Co 4	Co 5	Co 6

‘Sierra’ ‘Concerto’ Kombinationen, Woche 48

Si 1	**0.618**	0.054	0.043	0.196	-0.090	0.017
Si 2	-0.476	**0.018**	-0.041	-0.147	0.369	-0.028
Si 3	-0.020	0.044	**0.466**	-0.106	-0.020	-0.270
Si 4	0.184	0.239	0.012	**0.201**	-0.310	0.005
Si 5	0.122	-0.086	-0.109	-0.211	**0.321**	-0.012
Si 6	0.348	0.152	0.122	0.113	-0.269	**0.000**
	Co 1	Co 2	Co 3	Co 4	Co 5	Co 6

1 Gesamtbeurteilung, 2 Knospenbesatz, 3 Wurzelqualität, 4 Vergilbung, 5 Welke, 6 Krankheiten

Übersicht A9: Hauptkoordinatenanlyse der Spearman-Korrelationsmatrix für die Bonituren der Qualitätsmerkmale bei ‘Sierra’ und ‘Concerto’ in Woche 44 und 48

a) ‘Sierra’-‘Concerto’, Woche 44

*** Latent Roots ***						*** Percentage variation ***					
1	2	3	4	5	6	1	2	3	4	5	6
2.6314	2.1719	1.1077	1.0060	0.7696	0.5633	27.97	23.09	11.77	10.69	8.18	5.99
7	8	9	10	11	12	7	8	9	10	11	12
0.4897	0.2736	0.2070	0.1105	0.0771	0.0000	5.21	2.91	2.20	1.17	0.82	0.00

aus Hauptkoordinatenanalyse abgeleite Distanzmatrix

1	0.000										
2	1.351	0.000									
3	1.187	1.139	0.000								
4	1.148	1.574	1.417	0.000							
5	1.162	1.475	1.662	1.168	0.000						
6	1.314	1.674	1.737	1.381	0.953	0.000					
7	1.026	1.153	1.317	1.377	1.339	1.565	0.000				
8	1.180	1.376	1.248	1.506	1.499	1.456	0.781	0.000			
9	1.288	1.416	1.271	1.364	1.331	1.363	1.561	1.501	0.000		
10	1.139	1.355	1.644	1.369	1.170	1.346	0.903	1.195	1.427	0.000	
11	1.203	1.244	1.413	1.534	1.328	1.355	0.688	0.587	1.526	0.979	0.000
12	1.166	1.293	1.430	1.401	1.374	1.401	0.882	1.056	1.481	0.891	0.864
1	2	3	4	5	6	7	8	9	10	11	12

b) ‘Sierra’-‘Concerto’, Woche 48

*** Latent Roots ***						*** Percentage variation ***					
1	2	3	4	5	6	1	2	3	4	5	6
2.4088	1.9217	1.5342	1.0689	0.9725	0.8024	23.44	18.70	14.93	10.40	9.47	7.81
7	8	9	10	11	12	7	8	9	10	11	12
0.5535	0.4937	0.3137	0.1450	0.0601	0.0000	5.39	4.81	3.05	1.41	0.58	0.00

aus Hauptkoordinatenanalyse abgeleite Distanzmatrix

1	0.000										
2	1.422	0.000									
3	1.142	1.235	0.000								
4	1.105	1.466	1.328	0.000							
5	1.378	1.436	1.245	1.209	0.000						
6	0.948	1.517	1.340	0.966	1.112	0.000					
7	0.874	1.376	1.384	1.268	1.476	1.402	0.000				
8	1.718	1.401	1.443	1.515	1.124	1.434	1.717	0.000			
9	1.428	1.383	1.033	1.487	1.428	1.594	1.465	1.313	0.000		
10	1.277	1.234	1.406	1.264	1.619	1.411	1.086	1.568	1.412	0.000	
11	1.325	1.474	1.490	1.556	1.166	1.423	1.301	1.396	1.574	1.266	0.000
12	1.142	1.302	1.325	1.332	1.593	1.414	1.289	1.599	1.481	1.313	1.329
1	2	3	4	5	6	7	8	9	10	11	12

1 Gesamtbeurteilung, 2 Knospenbesatz, 3 Wurzelqualität, 4 Vergilbung, 5 Welke, 6 Krankheiten bei ‘Sierra’
7 Gesamtbeurteilung, 8 Knospenbesatz, 9 Wurzelqualität, 10 Vergilbung, 11 Welke, 12 Krankheiten bei ‘Concerto’

Übersicht A10: Spearman Rangkorrelationen der Substratanalysewerte in Variablenset 2

```
Spearman Rank Correlation

Sample size:                 19
Degrees of freedom =         17

Exact critical values for one-sided test:
 p=0.05, critical value = 0.399
 p=0.01, critical value = 0.564

*** Correlation matrix (adjusted for ties) ***

  1   1.000
  2   0.863   1.000
  3   0.874   0.818   1.000
  4  -0.164  -0.185  -0.108   1.000
  5   0.432   0.404   0.307   0.077   1.000
  6   0.574   0.606   0.650  -0.281   0.507   1.000
  7   0.489   0.540   0.504   0.221   0.763   0.563   1.000
  8  -0.341  -0.366  -0.110   0.284  -0.661  -0.360  -0.344   1.000
  9   0.014   0.212   0.092  -0.034   0.182   0.164   0.092  -0.206   1.000
 10   0.054   0.200   0.161   0.059   0.275   0.086   0.172  -0.333   0.519   1.000
 11   0.058   0.160   0.181  -0.165  -0.382   0.283  -0.142   0.129   0.345   0.174   1.000
 12   0.192   0.164   0.172  -0.216  -0.332  -0.011   0.026   0.511  -0.390  -0.507   0.090   1.000

          1       2       3       4       5       6       7       8       9      10      11      12

    Reihenfolge der Variablen:

    1 N, Woche 23           5 N, Woche 29           9 N, Woche 41
    2 K, Woche 23           6 K, Woche 29          10 K, Woche 41
    3 Salz, Woche 23        7 Salz, Woche 29       11 Salz, Woche 41
    4 pH-Wert, Woche 23     8 pH-Wert, Woche 29    12 pH-Wert, Woche 41
```

Übersicht A11: Normalverteilungstests der Substratanalysewerte in Variablenset 2 (Reihenfolge der Variablen wie in Übersicht A10)

Type of test	Variate(s)	Test statistic: Anderson-Darling	Cramer-von Mises	Watson
marginal				
	1	5.205 **	0.198 **	0.186 **
	2	4.794 **	0.089	0.089
	3	5.050 **	0.176 *	0.170 **
	4	5.017 **	0.087	0.084
	5	5.757 **	0.247 **	0.247 **
	6	5.927 **	0.334 **	0.299 **
	7	5.283 **	0.202 **	0.195 **
	8	5.095 **	0.088	0.081
	9	8.322 **	0.827 **	0.769 **
	10	7.048 **	0.598 **	0.548 **
	11	5.485 **	0.274 **	0.251 **
	12	5.169 **	0.102	0.091
bivariate angle				
1	2	3.975 **	0.135	0.139
1	3	4.272 **	0.038	0.014
1	4	4.632 **	0.041	0.052
1	5	4.008 **	0.060	0.066
1	6	4.213 **	0.055	0.061
1	7	4.279 **	0.052	0.041
1	8	4.242 **	0.166	0.159 ?
1	9	6.456 **	0.261	-0.027
1	10	6.060 **	0.234	-0.115
1	11	6.768 **	0.210	-0.130
1	12	4.212 **	0.141	0.145
2	3	4.296 **	0.014	-0.004
2	4	4.751 **	0.038	0.048
2	5	3.876 **	0.042	0.053
2	6	4.345 **	0.014	0.001
2	7	4.537 **	0.045	-0.018
2	8	4.287 **	0.124	0.129
2	9	7.221 **	0.290	-0.142
2	10	6.988 **	0.336	-0.139
2	11	6.954 **	0.214	-0.183
2	12	4.233 **	0.136	0.136
3	4	4.892 **	0.034	0.034
3	5	3.908 **	0.066	0.080
3	6	4.647 **	0.103	0.117
3	7	4.407 **	0.028	-0.040
3	8	4.396 **	0.155	0.151
3	9	6.934 **	0.283	-0.102
3	10	5.819 **	0.203	-0.096
3	11	7.577 **	0.259	-0.176
3	12	4.409 **	0.155	0.149
4	5	3.996 **	0.097	0.110
4	6	4.712 **	0.027	-0.043
4	7	4.727 **	0.017	-0.057
4	8	4.210 **	0.215	0.186 ?
4	9	6.085 **	0.180	-0.080
4	10	5.641 **	0.113	-0.047
4	11	4.769 **	0.054	-0.039
4	12	3.958 **	0.144	0.139
5	6	5.022 **	0.078	0.075
5	7	4.594 **	0.040	0.029
5	8	4.418 **	0.106	0.120
5	9	6.652 **	0.187	-0.001
5	10	6.934 **	0.228	-0.147
5	11	6.875 **	0.217	-0.159
5	12	4.466 **	0.116	0.128
6	7	4.704 **	0.080	-0.033
6	8	4.512 **	0.187	0.165 ?
6	9	7.364 **	0.325	-0.166
6	10	5.637 **	0.211	-0.056
6	11	7.132 **	0.265	-0.206
6	12	4.309 **	0.196	0.178 ?
7	8	4.305 **	0.127	0.134
7	9	6.411 **	0.208	-0.074
7	10	5.549 **	0.205	-0.037
7	11	6.471 **	0.190	-0.153
7	12	4.376 **	0.140	0.142
8	9	6.815 **	0.255	-0.136
8	10	5.566 **	0.121	-0.079
8	11	5.225 **	0.060	-0.083
8	12	4.266 **	0.284	0.240 *
9	10	5.562 **	0.340	0.219 *
9	11	5.280 **	0.176	-0.012
9	12	4.805 **	0.381 ?	0.345 **
10	11	4.979 **	0.116	0.058
10	12	4.612 **	0.259	0.248 *
11	12	4.318 **	0.269	0.229 *
radius		4.735 **	0.311	0.322 **

?, *, ** indicate significance at 10%, 5% and 1% levels respectively

Übersicht A12: Test auf multivariate Ausreißer in Variablenset 2

```
Total number of units in data =  20

  Number of units with complete data =  19
  No outliers were detected
  Number of iterations used =  1
  The threshold distance, D0, =  4.8783


   Unit_No      1      2      3      4      5      6      7      8      9     10
       Wts 1.0000 1.0000 1.0000 1.0000      * 1.0000 1.0000 1.0000 1.0000 1.0000
  Mahaldst  3.370  3.194  4.018  3.404      *  3.563  3.369  3.278  3.435  2.811

   Unit_No     11     12     13     14     15     16     17     18     19     20
       Wts 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
  Mahaldst  3.065  3.553  3.146  3.838  2.833  3.529  3.134  3.016  4.042  3.136

  Distances printed as missing are for units with incomplete data


Erklärung der Abkürzungen:

Unit_No:  Betriebsnummer
Wts:      Gewichtung nach CAMPBELL, 1980
Mahaldst: Mahalanobis-Distanz des Objekts zum Mittelwertsvektor (Zentroid)
```

Übersicht A13: Screeplot der Hauptkomponentenanalyse der Substratanalysewerte

No	Root	%%	Cum	%	Scree Diagram (* represents 2%)
1	4.9943	416	416	42	**********
2	2.1916	183	599	18	*********
3	1.6568	138	737	14	*******
4	1.2188	102	838	10	*****
5	0.7073	59	897	6	***
6	0.5706	48	945	5	***
7	0.2312	19	964	2	*
8	0.1888	16	980	2	*
9	0.0999	8	988	1	*
10	0.0755	6	995	1	*
11	0.0422	4	998	0	
12	0.0230	2	1000	0	

Scale: 1 asterisk represents 2 units.

No	Root	per-1000	Cum %%	Del1	Del2	Del3
1	4.9943	416	416	234	*	*
2	2.1916	183	599	45	189	181
3	1.6568	138	737	36	8	14
4	1.2188	102	838	43	-6	-37
5	0.7073	59	897	11	31	48
6	0.5706	48	945	28	-17	-42
7	0.2312	19	964	4	25	29
8	0.1888	16	980	7	-4	-9
9	0.0999	8	988	2	5	6
10	0.0755	6	995	3	-1	-2
11	0.0422	4	998	2	1	*
12	0.0230	2	1000	*	*	*

Übersicht A14a und b: Bestimmung der Anzahl ‘wesentlicher’ Hauptkomponenten nach VELICER, 1976 (a)) und EASTMENT & KRZANOWSKI, 1982 (b)) nach Hauptkomponentenanalyse der Substratanalysewerte

a) Velicer

Principal Components removed	fq-value
0	0.1736
1	0.1106
2	0.1322
3	0.1389
4	0.1770
5	0.1402
6	0.1665
7	0.1887
8	0.2541
9	0.3108
10	0.4730
11	1.0000

b) Eastment & Krzanowski

Principal Component	W-value
1	1.6277
2	0.0553
3	0.2738
4	0.9665
5	0.2395
6	0.8002
7	0.2226
8	0.1696
9	0.0559
10	0.0449
11	0.0124

Übersicht A15a und b: Hauptkomponenten-Residuen nach Hauptkomponentenanalyse der Substratanalysewerte und Betrachtung von einer Dimension (a)) beziehungsweise von zwei Dimensionen (b))

a) Hauptkomponenten-Residuen bei einer Hauptkomponente

Q-values (PCA residuals)

unitname	Q
1	3.436
2	3.225
3	15.874
4	7.388
6	6.308
7	4.689
8	3.367
9	12.258
10	5.568
11	4.727
12	11.167
13	8.813
14	7.666
15	2.539
16	2.251
17	4.426
18	5.376
19	11.418
20	5.606

Critical value at alpha 0.05000 for Q: 15.88

b) Hauptkomponenten-Residuen bei zwei Hauptkomponenten

Q-values (PCA residuals)

unitname	Q
1	3.350
2	1.266
3	10.570
4	7.054
6	5.469
7	2.995
8	2.188
9	2.921
10	5.460
11	3.647
12	11.108
13	2.709
14	7.438
15	1.756
16	1.829
17	4.354
18	2.943
19	5.084
20	4.514

Critical value at alpha 0.05000 for Q: 11.25

Übersicht A16: Approximation von Variablenwerten durch interaktives Vorgehen bei der Auswertung von Hauptkomponenten-Biplots mit Prediktionsmarkern

the predicted value for variable 4 and unit 9 is

1.012

unitname	salz23
1	0.810
2	0.770
3	1.820
4	0.700
6	1.250
7	1.390
8	1.540
9	**1.080**
10	1.490
11	1.720
12	0.730
13	0.780
14	1.830
15	1.030
16	1.010
17	0.890
18	0.720
19	1.300
20	0.950

The first two principal components explain 59.88 percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

N Woche 23	0.1455
K Woche 23	0.1566
pH Woche 23	0.0472
Salz Woche 23	**0.1693**
N Woche 29	0.2874
K Woche 29	0.1748
pH Woche 29	0.2488
Salz Woche 29	0.1372
N Woche 41	0.0933
K Woche 41	0.1067
pH Woche 41	0.1694
Salz Woche 41	0.2638

the predicted value for variable 2 and unit 14 is

366.2

unitname	k29
1	140.0
2	174.0
3	469.0
4	122.0
6	106.0
7	238.0
8	197.0
9	160.0
10	197.0
11	224.0
12	168.0
13	102.0
14	**377.0**
15	105.0
16	203.0
17	130.0
18	97.0
19	123.0
20	216.0

The first two principal components explain 59.88 percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

N Woche 23	0.1455
K Woche 23	0.1566
pH Woche 23	0.0472
Salz Woche 23	0.1693
N Woche 29	0.2874
K Woche 29	**0.1748**
pH Woche 29	0.2488
Salz Woche 29	0.1372
N Woche 41	0.0933
K Woche 41	0.1067
pH Woche 41	0.1694
Salz Woche 41	0.2638

the predicted value for variable 1 and unit 11 is

197.8

unitname	n41
1	123.0
2	31.0
3	906.0
4	24.0
6	9.0
7	26.0
8	140.0
9	24.0
10	55.0
11	**24.0**
12	324.0
13	159.0
14	100.0
15	42.0
16	14.0
17	26.0
18	41.0
19	53.0
20	80.0

The first two principal components explain 59.88 percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

N Woche 23	0.1455
K Woche 23	0.1566
pH Woche 23	0.0472
Salz Woche 23	0.1693
N Woche 29	0.2874
K Woche 29	0.1748
pH Woche 29	0.2488
Salz Woche 29	0.1372
N Woche 41	**0.0933**
K Woche 41	0.1067
pH Woche 41	0.1694
Salz Woche 41	0.2638

Übersicht A17: Hauptkomponentenanalyse der Schattiersollwerte

***** Principal components analysis *****

*** Latent Roots ***

1	2	3	4	5	6
26171	2402	768	263	142	122
7	8	9	10	11	12
67	36	11	6	0	0
13	14				
0	0				

*** Percentage variation ***

1	2	3	4	5	6
87.27	8.01	2.56	0.88	0.47	0.41
7	8	9	10	11	12
0.22	0.12	0.04	0.02	0.00	0.00
13	14				
0.00	0.00				

*** Trace ***

29987

*** Latent Vectors (Loadings) ***

	1	2	3	4	5
klx23	-0.26426	0.47357	0.33311	-0.39559	-0.34769
klx24	-0.26274	0.42957	0.14782	-0.15128	-0.07366
klx25	-0.26153	0.38748	-0.06949	0.13601	0.29530
klx26	-0.25441	0.33547	-0.15780	0.31623	0.48384
klx27	-0.24225	0.03370	-0.41399	0.28386	-0.32703
klx28	-0.24225	0.03370	-0.41399	0.28386	-0.32703
klx29	-0.23507	-0.14139	-0.26654	-0.26321	-0.11944
klx30	-0.23640	-0.13392	-0.26169	-0.34101	-0.06699
klx31	-0.23111	-0.19789	-0.13601	-0.30027	0.11422
klx32	-0.23111	-0.19789	-0.13601	-0.30027	0.11422
klx33	-0.27283	-0.19148	0.03169	-0.16470	0.52180
klx34	-0.32584	-0.18909	0.22279	0.21663	-0.06946
klx35	-0.32562	-0.26228	0.36573	0.22474	-0.09098
klx36_42	-0.32562	-0.26228	0.36573	0.22474	-0.09098

	6	7	8	9	10
klx23	-0.18420	-0.11122	-0.10071	-0.09503	-0.03953
klx24	-0.01816	-0.00393	0.03586	0.00692	-0.12856
klx25	0.21668	0.13795	0.27258	0.31844	0.64202
klx26	0.23515	-0.02387	-0.17976	-0.27921	-0.51099
klx27	-0.22603	-0.16921	0.00114	0.03217	0.02260
klx28	-0.22603	-0.16921	0.00114	0.03217	0.02260
klx29	0.13774	0.45936	0.31175	-0.65478	0.11932
klx30	0.16418	0.34938	0.09575	0.61470	-0.43701
klx31	0.22613	-0.38897	-0.23782	-0.02931	0.15391
klx32	0.22613	-0.38897	-0.23782	-0.02931	0.15391
klx33	-0.75244	-0.00376	0.14898	0.00581	-0.00883
klx34	-0.06061	0.48940	-0.68758	0.03138	0.19321
klx35	0.14709	-0.13332	0.28941	0.00891	-0.09376
klx36_42	0.14709	-0.13332	0.28941	0.00891	-0.09376

	11	12	13	14
klx23	-0.50052	0.00000	0.00000	0.00000
klx24	0.82345	0.00000	0.00000	0.00000
klx25	-0.13059	0.00000	0.00000	0.00000
klx26	-0.19101	0.00000	0.00000	0.00000
klx27	-0.00025	-0.08399	0.70210	0.00027
klx28	-0.00025	0.08399	-0.70210	-0.00027
klx29	0.00335	0.00000	0.00000	0.00000
klx30	-0.09950	0.00000	0.00000	0.00000
klx31	0.04673	-0.70209	-0.08399	-0.00296
klx32	0.04673	0.70209	0.08399	0.00296
klx33	-0.00096	0.00000	0.00000	0.00000
klx34	0.04911	0.00000	0.00000	0.00000
klx35	-0.02428	-0.00291	-0.00062	0.70710
klx36_42	-0.02428	0.00291	0.00062	-0.70710

Übersicht A18: Variablen und verwendete Proximitätsmaße im Variablenset 3

Variable	Proximitätsmaß	Formel zur Berechnung der Ünähnlichkeit von zwei Objekten r und t
erste Hauptkomponente der Schattiersollwerte	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
zweite Hauptkomponente der Schattiersollwerte	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Schattierfabe (ja/nein)	jaccard	wenn $x_r = x_t = 1$, dann 1 $x_r = x_t = 0$, dann 0 $x_r \neq x_t$, dann 0
Anzahl Rückvorgänge	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
aufgestellt mit ... Pflanzen	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
Endstand	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
Wochen auf Endstand	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
längste Phase mit einer Standweite	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
Verhältnis Aufstellen : Endstand	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Verhältnis Stand vor und nach längster Phase	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Nettowochenquadratmeter	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
negative DIF Juni	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
negative DIF Rest	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Tagesmitteltemperatur Juni	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Tagesmitteltemperatur Rest	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Lüftungstemperatur Juni	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Lüftungstemperatur Rest	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$
Lüftung über Heizung Juni	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
Lüftung über Heizung Rest	cityblock	$1-\left\|x_r-x_t\right\|/\text{Spannweite}$
geschätzeter Energieverbrauch	euklidisch	$1-\left(\left(x_r-x_t\right)/\text{Spannweite}\right)^2$

r und t bezeichnen zwei Objekte, x_r, ist also der Wert von Variable x bei Objekt r

Übersicht A19: Hauptkoordinatenanalyse und ordinale mehrdimensionale Skalierung von Variablenset 3

```
*****  Principal coordinates analysis  *****

 ***  Latent Roots  ***

    pco3['Roots']
            1           2           3           4           5           6
       0.9128      0.6405      0.4746      0.4438      0.3993      0.3067
            7           8           9          10          11          12
       0.2403      0.1960      0.1746      0.1403      0.0999      0.0706
           13          14          15          16          17          18
       0.0561      0.0513      0.0292      0.0243      0.0133      0.0118
           19          20
       0.0000     -0.0016

 ***  Percentage variation  ***

    pco3['Roots']
            1           2           3           4           5           6
        21.31       14.95       11.08       10.36        9.32        7.16
            7           8           9          10          11          12
         5.61        4.58        4.08        3.27        2.33        1.65
           13          14          15          16          17          18
         1.31        1.20        0.68        0.57        0.31        0.27
           19          20
         0.00       -0.04

 ***  Trace  ***

  pco3['Trace']
        4.284

 *  Some roots are negative - non-Euclidean distance matrix  *

-------------------------------------------------------------------------------------------------

 *****  Multidimensional scaling  *****

 ***  Least-squares scaling criterion  ***
 *  Distances fitted using monotonic regression (non-metric MDS)  *
 *  Primary treatment of ties  *


*** Stress         0.2021 (two dimensions fitted)
*** Stress         0.1260 (three dimensions fitted)
*** Stress         0.0851 (four dimensions fitted)

 ***  Coordinates  ***
                    mds3_4
                         1           2           3           4
             1      0.5659      0.6257      0.2294     -0.3262
             2      0.9403      0.0201     -0.3504     -0.0147
             3     -0.0100     -0.4604     -0.0666      0.8758
             4      0.8986      0.1208      0.2140     -0.1418
             5      0.0380     -0.4508     -0.2937      0.0391
             6     -0.0084     -0.2376      0.1034     -0.0272
             7      1.2147     -0.3552      0.4728      0.2115
             8     -0.2717      1.2454      0.5188      0.0254
             9     -0.3283      0.1616     -0.1366      0.2875
            10     -0.5788     -0.7478      0.6278      0.5046
            11     -0.6459      0.1144     -0.1809     -0.5027
            12     -0.3325     -0.0886     -1.1563     -0.1067
            13     -0.1188     -0.0687      0.1565     -0.3042
            14     -1.2610      0.7064     -0.3287      0.1097
            15     -0.7132      0.1074      0.7053      0.2898
            16     -0.2396     -0.0181      0.5698     -0.5013
            17      0.1706     -0.4671      0.0480     -0.3611
            18     -0.5113     -1.0390     -0.2358     -0.4657
            19      0.4623      0.5225     -0.6423      0.6604
            20      0.7291      0.3089     -0.2545     -0.2521
```

Übersicht A20: Nächste Nachbarn, typische Objekte und Ähnlichkeit zwischen Gruppierungen im Variablenset 3

**** Neighbours table derived from sim3 ****

1	4	90.5	20	88.9	19	87.6	2	85.7	13	82.6
2	20	94.5	4	91.2	1	85.7	6	85.0	7	84.3
3	6	82.3	5	81.4	10	80.3	13	80.2	19	80.1
4	2	91.2	20	91.2	1	90.5	7	88.7	6	85.1
5	6	92.2	13	91.3	9	90.6	17	90.1	20	84.5
6	17	95.4	13	93.4	5	92.2	9	91.2	16	88.2
7	4	88.7	20	87.5	2	84.3	17	79.3	3	78.6
8	14	80.0	6	76.6	9	76.2	20	75.9	11	75.3
9	6	91.2	5	90.6	13	89.8	17	85.4	16	82.9
10	6	84.0	15	81.6	3	80.3	17	78.2	18	77.6
11	16	86.4	18	85.1	13	83.5	12	83.4	17	82.5
12	11	83.4	6	80.9	5	79.8	20	78.5	2	77.9
13	6	93.4	16	93.3	5	91.3	9	89.8	17	89.6
14	11	81.6	8	80.0	12	77.0	9	76.6	15	75.6
15	13	83.3	6	82.8	16	82.6	9	82.6	11	82.4
16	13	93.3	6	88.2	11	86.4	17	84.6	9	82.9
17	6	95.4	5	90.1	13	89.6	20	87.0	9	85.4
18	6	86.1	13	85.5	11	85.1	5	83.6	17	83.5
19	1	87.6	20	82.2	3	80.1	5	80.0	2	79.3
20	2	94.5	4	91.2	1	88.9	7	87.5	17	87.0

**** Mean similarities between and within groups ****
**** Similarity matrix: sim3 ****

** Between and within groups similarity matrix **

1	71.4				
2	72.6	81.4			
3	73.3	82.5	86.4		
4	66.8	77.4	80.9	77.6	
5	73.3	76.3	80.1	67.9	87.0

1	2	3	4	5

**** Most typical members ****
**** Similarity matrix: sim3 ****

group 1		group 2		group 3		group 4		group 5	
8	76.2	13	85.2	6	89.9	18	77.6	20	89.2
14	70.9	16	83.6	5	87.9	10	77.6	1	88.2
7	67.1	11	82.9	17	87.6			4	87.9
		9	82.1	3	80.3			2	87.7
		15	79.6					19	82.0
		12	75.2						

Übersicht A21: Multiple Korrespondenzanalyse der betrieblichen Strukturdaten in Variablenset 4

Squared singular values

	eigenv	%Roots	Cum%Root
1	0.25528	25.53	25.53
2	0.22899	22.90	48.43
3	0.18842	18.84	67.27
4	0.11500	11.50	78.77
5	0.08902	8.90	87.67
6	0.04986	4.99	92.66
7	0.03574	3.57	96.23
8	0.02574	2.57	98.80
9	0.01196	1.20	100.00
10	0.00000	0.00	100.00
11	0.00000	0.00	100.00
12	0.00000	0.00	100.00
13	0.00000	0.00	100.00
14	0.00000	0.00	100.00
15	0.00000	0.00	100.00
16	0.00000	0.00	100.00
17	0.00000	0.00	100.00

Normalkoordinaten Betriebe

betrieb	rowvar[1]	rowvar[2]
1.00	-0.5790	-0.2266
2.00	-0.7717	0.0992
3.00	**1.0115**	**-0.8311**
4.00	-0.4538	0.1301
5.00	-0.6980	-0.1119
6.00	0.0252	-0.1852
7.00	0.1588	0.4969
8.00	0.5614	-0.2028
9.00	-0.1724	0.2621
10.00	0.1949	-0.7023
11.00	0.1249	0.6840
12.00	-0.6249	-0.0998
13.00	0.0256	0.6178
14.00	0.3434	0.6488
15.00	-0.2223	-0.7994
16.00	0.5362	0.3229
17.00	-0.2223	-0.7994
18.00	0.8454	0.0153
19.00	0.4172	0.4376
20.00	-0.5000	0.2438

Überprüfung der Interpolation

xline	yline
-0.402	**-0.268**
1.381	**0.900**
1.534	**1.178**
1.573	**0.153**
3.962	**-3.630**
1.228	**-2.187**
1.041	**-1.760**
-0.127	**-1.373**
-1.086	**-0.494**

mean of xline and yline

mxline	myline
1.012	**-0.8312**

Benennungen **(Betrieb 3 fett)**

Nummer	Abkürzung	Beschreibung
1	**ee**	**einheitserden**
2	vm1	ein vermarktungsweg
3	**fg10**	**ueber 10000 qm**
4	mg50	ueber 50000 stck
5	sf1_m	stellflaeche 1 modern
6	sf2_m	stellflaeche 2 modern
7	**bw1_k**	**bewaesserung 1 ueber kopf**
8	**bw2_k**	**bewaesserung 2 ueber kopf**
9	**wes**	**westliches muensterland**
10	subs	andere substrate
11	**vmg1**	**mehrere vermarktungswege**
12	fw10	unter 10000 qm
13	**mw50**	**unter 50000 stck**
14	**sf1_a**	**stellflaeche 1 traditionell**
15	**sf2_a**	**stellflaeche 2 traditionell**
16	bw1_f	bewaesserung 1 von unten
17	bw2_f	bewaesserung 2 von unten
18	ost	oestliches muensterland

Normalkoordinaten Merkmale / Standardkoordinaten Merkmale

	colvar[1]	colvar[2]	scolvar[1]	scolvar[2]
1	-0.2033	-0.1283	**-0.402**	**-0.268**
2	0.0524	0.5375	0.104	1.123
3	0.5259	-0.8420	**1.041**	**-1.760**
4	-0.3340	0.5634	-0.661	1.177
5	-0.8526	-0.5262	-1.688	-1.100
6	-0.6504	-0.0599	-1.287	-0.125
7	0.7748	0.5636	**1.534**	**1.178**
8	2.0019	-1.7368	**3.962**	**-3.630**
9	-0.5487	-0.2362	**-1.086**	**-0.494**
10	0.2485	0.1568	0.492	0.328
11	-0.0641	-0.6570	**-0.127**	**-1.373**
12	-0.3506	0.5614	-0.694	1.173
13	0.6203	-1.0464	**1.228**	**-2.187**
14	0.6976	0.4306	**1.381**	**0.900**
15	0.7950	0.0732	**1.573**	**0.153**
16	-0.5165	-0.3757	-1.022	-0.785
17	-0.1054	0.0914	-0.209	0.191
18	0.2352	0.1012	0.465	0.212

Übersicht A22: Vorhersage (Prediktion) der Klassenzugehörigkeit durch die Prediktionsregionen im multiplen Korrespondenzanalyse-Biplot bei Verwendung der Chi-Quadrat-Distanz (mca) und des extended matching-Koeffizienten (emc), sowie Beschreibung der Klassen und wahre Klassenhäufigkeiten

	Wahre Merkmalsklasse je Variable und Merkmalsklasse nach Prediktion durch verwendetes Distanzmaß und Distanzmaß																			
Betrieb	**menge**	**mca**	**emc**	**groesse**	**mca**	**emc**	**absatz**	**mca**	**emc**	**region**	**mca**	**emc**	**stellf**	**mca**	**emc**	**be-wäs**	**mca**	**emc**	**subst**	**mca**	**emc**
1	2	2	2	2	**1**	**1**	2	**1**	**1**	2	**1**	2	2	2	2	1	1	1	2	2	2
2	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2	1	1	1	2	2	2
3	1	1	1	2	2	2	1	1	1	2	**1**	**1**	1	1	1	2	2	2	2	**1**	**1**
4	2	2	2	1	1	1	2	2	2	2	**1**	2	2	2	2	1	1	1	2	2	2
5	2	2	2	1	1	1	1	**2**	1	2	2	2	2	2	2	1	1	1	1	**2**	**2**
6	1	**2**	1	1	1	**2**	1	**2**	1	1	1	1	1	1	**2**	1	1	1	1	**2**	1
7	2	2	2	1	1	1	2	2	2	1	1	1	1	1	1	1	**2**	**2**	1	1	1
8	1	1	1	2	2	2	2	2	**1**	1	1	1	1	1	1	1	**2**	**2**	1	1	1
9	2	2	2	1	1	1	1	**2**	**2**	2	2	2	1	1	1	2	2	2	2	2	2
10	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	1	1	1	1	**2**	1
11	2	2	2	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	1	1	**2**
12	2	2	2	1	1	1	1	**2**	1	1	1	**2**	2	2	2	1	1	1	2	2	2
13	2	2	2	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	2	**1**	2
14	2	2	2	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	2	**1**	**1**
15	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	1	1	1	2	2	2
16	2	2	2	2	**1**	**1**	2	2	2	1	1	1	1	1	1	2	2	2	2	**1**	**1**
17	1	1	1	2	2	2	1	1	1	1	1	1	2	2	2	1	1	1	2	2	2
18	1	1	1	2	2	2	2	2	2	1	1	1	1	1	1	2	2	2	1	1	1
19	2	2	2	1	1	1	1	**2**	**2**	1	1	1	1	1	1	2	2	2	1	1	1
20	2	2	2	1	1	1	2	2	2	1	1	**2**	2	2	2	1	1	1	1	**2**	**2**
Fehler	**-**	**1**	**0**	**-**	**2**	**3**	**-**	**6**	**4**	**-**	**3**	**3**	**-**	**0**	**1**	**-**	**2**	**2**	**-**	**8**	**6**

Count menge 1.00 (mw50) 7 2.00 (mg50 13 Count groesse 1.00 (fw10) 12 2.00 (fg10) 8	Count absatzwege 1.00 (vmg1) 9 2.00 (vm1) 11 Count region 1.00 (ost) 14 2.00 (west) 6	Count stellflaeche 1.00 (sf1_a) 11 2.00 (sf1_m) 9 Count bewässerung 1.00 (bw1_f) 12 2.00 (bw1_k) 8 Count substrate 1.00 (subs) 9 2.00 (ee) 11

Übersicht A23: Hauptkoordinatenanalyse, ordinale, mehrdimensionale Skalierung, nächste Nachbarn und Zentroid Distanzen; Grundlagen für die Abbildungen A59, A60 und A61

***** Principal coordinates analysis *****

*** Latent Roots ***

1	2	3	4	5	6
0.4990	0.4743	0.4086	0.4060	0.3840	0.3162
7	8	9			
0.2986	0.2650	0.0000			

*** Percentage variation ***

1	2	3	4	5	6
16.35	15.54	13.39	13.30	12.58	10.36
7	8	9			
9.78	8.69	0.00			

*** Centroid distances ***

1	2	3	4	5
0.5433	0.5671	0.5774	0.6408	0.5512
6	7	8	9	
0.5880	0.6128	0.5910	0.5623	

***** Multidimensional scaling *****

*** Least-squares scaling criterion ***
* Distances fitted using monotonic regression (non-metric MDS) *
* Primary treatment of ties *

*** Stress 0.1220

Koordinaten der ersten und zweiten Dimension der Hauptkoordinatenanalyse

	1	2
1	-0.1391	-0.0677
2	-0.0035	-0.0998
3	-0.1891	-0.2743
4	-0.1491	0.5746
5	-0.2337	-0.1876
6	0.3865	0.0583
7	0.4523	-0.0912
8	-0.1139	0.0017
9	-0.0105	0.0860

Koordinaten der ersten und zweiten Dimension der ordinalen mehrdimensionalen Slkalierung

	1	1
1	0.0583	-0.0277
2	-0.6503	-0.2442
3	-0.3891	-0.9689
4	1.6076	0.4992
5	0.1200	-0.5295
6	-0.3239	1.0031
7	-1.3279	0.3540
8	0.9030	-0.5307
9	0.0023	0.4448

Proximitätsmatrix der Residuen (padist)

1	0.0000								
2	0.6340	0.0000							
3	0.7037	0.7512	0.0000						
4	0.7997	0.8427	0.8775	0.0000					
5	0.6384	0.7086	0.6001	0.8232	0.0000				
6	0.7491	0.7852	0.8459	0.8773	0.7736	0.0000			
7	0.8029	0.7465	0.7809	0.9042	0.8235	0.6880	0.0000		
8	0.7036	0.7517	0.7857	0.8395	0.7051	0.7779	0.8471	0.0000	
9	0.6773	0.7265	0.7072	0.7832	0.7139	0.6669	0.8380	0.7846	0.0000
	1	2	3	4	5	6	7	8	9

**** Neighbours table derived from simdis (simdis=1-padist/max(padist) ****

1	2	29.9	5	29.4	9	25.1	8	22.2	3	22.2	1 - struktur
2	1	29.9	5	21.6	9	19.6	7	17.4	3	16.9	2 - platzbedarf
3	5	33.6	1	22.2	9	21.8	2	16.9	7	13.6	3 - temperatur
4	9	13.4	1	11.6	5	8.9	8	7.2	2	6.8	4 - schattierung
5	3	33.6	1	29.4	8	22.0	2	21.6	9	21.0	5 - substrate
6	9	26.2	7	23.9	1	17.2	5	14.4	8	14.0	6 - Concerto 44
7	6	23.9	2	17.4	3	13.6	1	11.2	5	8.9	7 - Concerto 48
8	1	22.2	5	22.0	2	16.9	6	14.0	9	13.2	8 - Sierra 44
9	6	26.2	1	25.1	3	21.8	5	21.0	2	19.6	9 - Sierra 48

Übersicht A24: Variablensets für die generalisierte kanonische Analyse

Benennung in Übersichten	Benennung in Abbil-dungen	Beschreibung	Rangtrans-formation	Meßniveau
BIN_ABS	absw	Absatzwege	nein	single nominal
BIN_BEW1	bew1	Bewässerung 1	nein	single nominal
BIN_BEW2	bew2	Bewässerung 2	nein	single nominal
BIN_GROE	groe	Betriebsgröße	nein	single nominal
BIN_KREI	krei	Region	nein	single nominal
BIN_MEN	men	Produktionsumfang	nein	single nominal
BIN_SF1	sf1	Stellfläche 1	nein	single nominal
BIN_SF2	sf2	Stellfläche 2	nein	single nominal
BIN_SUBS	subs	Substrate	nein	single nominal
ANZ_RUE	rück	Anzahl Rückvorgänge	nein	single nominal
AUFS_END	a:e	Aufstellen : Rücken	nein	ordinal
AUFSTELL	auf	Aufstellen mit ...	nein	ordinal
ENDSTAND	end	Endstand mit ...	nein	ordinal
MAX_PER	maxper	längste Periode	nein	ordinal
MAX_RED	maxred	größte Reduktion	nein	ordinal
NET_JAQM	flae	Nettojahres-qm	nein	numerical
WOAUFEND	woend	Wochen im Endstand	nein	ordinal
DIF_JUN	dif1	Dif im Juni	nein	single nominal
DIF_RES	dif2	Dif Restzeit	nein	single nominal
LUFT_JUN	luft1	Lüftung im Juni	nein	ordinal
LUFT_RES	luft2	Lüftung Restzeit	nein	ordinal
TMT_JUN	tmt1	Tagesmitteltemp. Juni	nein	numerical
TMT_REST	tmt2	Tagesmitteltemp. Rest	nein	numerical
RENERGIE	ener	Energieverbrauch	ja	numerical
RLUGTEJ	lugte1	Lüftung>Heizung Juni	ja	ordinal
RLUGTER	lugte2	Lüftung>Heizung Rest	ja	ordinal
RVKLX1	licht1	1. Hk-Werte Schatten	ja	ordinal
RVKLX2	licht2	2. Hk-Werte Schatten	ja	ordinal
RFARBE	farbe	Schattierfarbe	nein	single nominal
RK23	k23	Kalium Woche 23	ja	ordinal
RK29	k29	Kalium Woche 29	ja	ordinal
RK41	k41	Kalium Woche 41	ja	ordinal
RN23	n23	Stickstoff Woche 23	ja	ordinal
RN29	n29	Stickstoff Woche 29	ja	ordinal
RN41	n41	Stickstoff Woche 41	ja	ordinal
RPH23	ph23	pH-Wert Woche 23	ja	ordinal
RPH29	ph29	pH-Wert Woche 29	ja	ordinal
RPH41	ph41	pH-Wert Woche 41	ja	ordinal
RSALZ23	sal23	Salzgehalt Woche 23	ja	ordinal
RSALZ29	sal29	Salzgehalt Woche 23	ja	ordinal
RSALZ41	sal41	Salzgehalt Woche 23	ja	ordinal
S_GES_44	S44ges	Gesamtbeurteilung	nein	ordinal
S_GES_48	S48ges	'Sierra' (S)		
C_GES_44	C44ges	'Concerto' (C)		
C_GES_48	C48ges	Woche 44 ode r 48		
S_GIL_44	S44gil	Vergilbung	nein	ordinal
S_GIL_48	S48gil	'Sierra' (S)		
C_GIL_44	C44gil	'Concerto' (C)		
C_GIL_48	C48gil	Woche 44 oder 48		
S_KNO_44	S44kno	Knospenbesatz	nein	ordinal
S_KNO_48	S48kno	'Sierra' (S)		
C_KNO_44	C44kno	'Concerto' (C)		
C_KNO_48	C48kno	Woche 44 oder 48		
S_KRA_44	S44kra	Krankheitsbefall	nein	ordinal
S_KRA_48	S48kra	'Sierra' (S)		
C_KRA_44	C44kra	'Concerto' (C)		
C_KRA_48	C48kra	Woche 44 oder 48		
S_WEL_44	S44wel	Welke	nein	ordinal
S_WEL_48	S48wel	'Sierra' (S)		
C_WEL_44	C44wel	'Concerto' (C)		
C_WEL_48	C48wel	Woche 44 oder 48		
S_WUR_44	S44wur	Wurzelqualität	nein	ordinal
S_WUR_48	S48wur	'Sierra' (S)		
C_WUR_44	C44wur	'Concerto' (C)		
C_WUR_48	C48wur	Woche 44 oder 48		

Übersicht A25: Loss-Werte der vier generalisierten kanonischen Analysen in den ersten beiden Dimensionen

'Sierra' Woche 44 als Set 6

Summary of Analysis

Loss per Set

	Sum	Dimension 1	2
Set 1 (struktur)	,347	,051	,296
Set 2 (platzbedarf)	,012	,001	,011
Set 3 (temperatur)	,024	,001	,022
Set 4 (schattierung)	,148	,051	,097
Set 5 (substrate)	,000	,000	,000
Set 6 ('Sierra' 44)	,096	,039	,057
Mean	,105	,024	,081
Fit	1,895		
Eigenvalue		,976	,919

'Sierra' Woche 48 als Set 6

Summary of Analysis

Loss per Set

	Sum	Dimension 1	2
Set 1 (struktur)	,125	,026	,099
Set 2 (platzbedarf)	,001	,001	,000
Set 3 (temperatur)	,012	,003	,009
Set 4 (schattierung)	,093	,019	,074
Set 5 (substrate)	,000	,000	,000
Set 6 ('Sierra' 48)	,031	,005	,026
Mean	,044	,009	,035
Fit	1,956		
Eigenvalue		,991	,965

'Concerto' Woche 44 als Set 6

Summary of Analysis

Loss per Set

	Sum	Dimension 1	2
Set 1 (struktur)	,287	,047	,240
Set 2 (platzbedarf)	,000	,000	,000
Set 3 (temperatur)	,024	,020	,004
Set 4 (schattierung)	,093	,035	,059
Set 5 (substrate)	,000	,000	,000
Set 6 ('Concerto' 44)	,096	,026	,069
Mean	,083	,021	,062
Fit	1,917		
Eigenvalue		,979	,938

'Concerto' Woche 48 als Set 6

Summary of Analysis

Loss per Set

	Sum	Dimension 1	2
Set 1 (struktur)	,272	,034	,238
Set 2 (platzbedarf)	,001	,001	,000
Set 3 (temperatur)	,027	,012	,015
Set 4 (schattierung)	,191	,061	,130
Set 5 (substrate)	,000	,000	,000
Set 6 ('Concerto' 48)	,205	,071	,134
Mean	,116	,030	,086
Fit	1,884		
Eigenvalue		,970	,914

Übersicht A26: Multiple Anpassungswerte der Variablensets nach generalisierter kanonischer Analyse, ‘Sierra’

a) ‘Sierra’ Woche 44 in Variablenset 6

Multiple Fit

	Dimension		
	Sum	1	2
BIN_ABS	,104	,084	,019
BIN_BEW1	1,584	,222	1,363
BIN_BEW2	,433	,432	,001
BIN_GROE	,011	,010	,001
BIN_KREI	,102	,056	,046
BIN_MEN	,081	,061	,020
BIN_SF1	,979	,003	,977
BIN_SF2	,069	,023	,046
BIN_SUBS	,257	,161	,095
ANZ_RUE	,281	,166	,114
AUFS_END	,414	,288	,126
AUFSTELL	,410	,298	,111
ENDSTAND	3,162	3,099	,063
MAX_PER	,106	,084	,022
MAX_RED	,416	,047	,369
NET_JAQM	1,533	1,388	,145
WOAUFEND	,460	,336	,123
DIF_JUN	,055	,021	,033
DIF_REST	,172	,052	,120
LUFT_JUN	1,091	,366	,724
LUFT_RES	4,284	4,043	,241
TMT_JUN	,727	,359	,368
TMT_REST	1,271	,923	,349
RENERGIE	,350	,153	,197
RLUGTEJ	,606	,591	,015
RLUGTER	,723	,718	,005
RVKLX1	1,437	1,259	,178
RVKLX2	,960	,136	,825
RFARBE	,239	,211	,027
RK23	,698	,528	,170
RK29	,226	,016	,210
RK41	,747	,101	,646
RN23	,433	,275	,157
RN29	,168	,100	,068
RN41	,059	,017	,042
RPH23	,228	,179	,049
RPH29	,368	,163	,206
RPH41	,545	,355	,190
RSALZ23	,390	,044	,346
RSALZ29	,074	,008	,066
RSALZ41	,215	,213	,002
S_GES_44	,731	,709	,022
S_GIL_44	,215	,006	,209
S_KNO_44	,432	,170	,262
S_KRA_44	,620	,380	,239
S_WEL_44	1,601	,001	1,600
S_WUR_44	2,135	2,121	,014

b) ‘Sierra’ Woche 48 in Variablenset 6

Multiple Fit

	Dimension		
	Sum	1	2
BIN_ABS	,130	,115	,015
BIN_BEW1	,089	,051	,038
BIN_BEW2	,497	,453	,043
BIN_GROE	,097	,003	,094
BIN_KREI	,501	,338	,163
BIN_MEN	1,010	,977	,032
BIN_SF1	,386	,180	,206
BIN_SF2	,184	,139	,045
BIN_SUBS	,563	,106	,457
ANZ_RUE	,878	,574	,304
AUFS_END	,809	,032	,777
AUFSTELL	,172	,084	,087
ENDSTAND	1,783	1,758	,025
MAX_PER	,363	,045	,318
MAX_RED	,232	,081	,150
NET_JAQM	2,952	2,939	,014
WOAUFEND	1,635	,256	1,379
DIF_JUN	,484	,404	,080
DIF_REST	,744	,001	,743
LUFT_JUN	,261	,172	,090
LUFT_RES	,332	,006	,327
TMT_JUN	,014	,003	,011
TMT_REST	,774	,037	,738
RENERGIE	,233	,157	,077
RLUGTEJ	1,534	,980	,554
RLUGTER	,706	,689	,017
RVKLX1	1,986	1,223	,763
RVKLX2	,835	,662	,173
RFARBE	1,057	,017	1,040
RK23	1,007	1,004	,003
RK29	,849	,262	,587
RK41	,479	,068	,411
RN23	,603	,294	,308
RN29	,277	,250	,028
RN41	,182	,163	,018
RPH23	,245	,240	,005
RPH29	,508	,001	,506
RPH41	,190	,075	,115
RSALZ23	,618	,027	,591
RSALZ29	,045	,001	,044
RSALZ41	,817	,703	,113
S_GES_48	1,574	,949	,626
S_GIL_48	,583	,000	,583
S_KNO_48	,877	,417	,460
S_KRA_48	1,928	1,916	,012
S_WEL_48	,060	,041	,019
S_WUR_48	,377	,024	,353

Übersicht A27: Multiple Anpassungswerte der Variablensets nach generalisierter kanonischer Analyse, ‘Concerto’

a) ‘Concerto’ Woche 44 in Variablenset 6

Multiple Fit

	Dimension		
	Sum	1	2
BIN_ABS	,263	,150	,113
BIN_BEW1	,936	,285	,651
BIN_BEW2	,708	,702	,006
BIN_GROE	,093	,002	,092
BIN_KREI	,049	,018	,031
BIN_MEN	,924	,856	,069
BIN_SF1	1,088	,319	,769
BIN_SF2	,153	,152	,001
BIN_SUBS	,063	,048	,014
ANZ_RUE	,327	,015	,312
AUFS_END	,392	,025	,367
AUFSTELL	,588	,266	,322
ENDSTAND	1,682	,490	1,192
MAX_PER	,219	,179	,040
MAX_RED	,373	,004	,369
NET_JAQM	,629	,620	,009
WOAUFEND	,691	,579	,111
DIF_JUN	,699	,189	,510
DIF_REST	,997	,102	,895
LUFT_JUN	1,319	,478	,840
LUFT_RES	1,178	,444	,734
TMT_JUN	,009	,007	,002
TMT_REST	,279	,020	,259
RENERGIE	,459	,282	,177
RLUGTEJ	,835	,384	,451
RLUGTER	,359	,222	,137
RVKLX1	1,674	1,593	,080
RVKLX2	1,121	,286	,835
RFARBE	,453	,355	,098
RK23	,647	,434	,212
RK29	,067	,003	,064
RK41	,985	,454	,531
RN23	,470	,062	,407
RN29	,456	,127	,328
RN41	,333	,067	,266
RPH23	,172	,171	,001
RPH29	,454	,093	,362
RPH41	,142	,084	,058
RSALZ23	,498	,134	,364
RSALZ29	,124	,001	,123
RSALZ41	,241	,239	,002
C_GES_44	2,495	2,271	,224
C_GIL_44	,229	,214	,015
C_KNO_44	1,683	,382	1,302
C_KRA_44	,433	,101	,332
C_WEL_44	,096	,084	,013
C_WUR_44	,330	,309	,021

b) ‘Concerto’ Woche 48 in Variablenset 6

Multiple Fit

	Dimension		
	Sum	1	2
BIN_ABS	,706	,207	,499
BIN_BEW1	1,080	,152	,928
BIN_BEW2	,258	,213	,045
BIN_GROE	,012	,011	,001
BIN_KREI	,504	,452	,052
BIN_MEN	,292	,267	,025
BIN_SF1	1,595	1,367	,228
BIN_SF2	,520	,039	,481
BIN_SUBS	,639	,069	,570
ANZ_RUE	1,269	,004	1,265
AUFS_END	,129	,101	,027
AUFSTELL	,174	,019	,155
ENDSTAND	1,415	,128	1,288
MAX_PER	,154	,145	,009
MAX_RED	,429	,384	,045
NET_JAQM	,696	,366	,329
WOAUFEND	,488	,334	,153
DIF_JUN	,428	,023	,405
DIF_REST	,386	,001	,385
LUFT_JUN	,882	,019	,863
LUFT_RES	2,051	1,981	,069
TMT_JUN	2,751	1,897	,855
TMT_REST	2,153	1,727	,427
RENERGIE	,924	,292	,633
RLUGTEJ	,674	,538	,136
RLUGTER	,064	,051	,013
RVKLX1	1,584	1,458	,126
RVKLX2	,385	,081	,304
RFARBE	,726	,724	,001
RK23	,830	,628	,202
RK29	,773	,747	,026
RK41	1,306	,103	1,203
RN23	,628	,420	,208
RN29	,126	,027	,099
RN41	,336	,336	,000
RPH23	,073	,008	,065
RPH29	,645	,269	,376
RPH41	,344	,334	,010
RSALZ23	,841	,341	,500
RSALZ29	,166	,010	,157
RSALZ41	,271	,033	,238
C_GES_48	,259	,254	,005
C_GIL_48	,219	,172	,046
C_KNO_48	,343	,048	,295
C_KRA_48	,646	,572	,074
C_WEL_48	,284	,221	,063
C_WUR_48	1,132	,038	1,093

Anhang Teil II B

Übersichten zur Auswertung der Kennzahlen der Topfpflanzenbetriebe 1992 bis 1994, Kapitel 3.2

Übersicht B1: Ausgewählte Kennzahlen und Gruppierungskriterien

Gruppierungsdaten

abswg	Absatzweg (indirekt, direkt, indirekt und direkt)
fak	Anzahl Arbeitskräfte (<=2 AK, >2 - <=5 AK, >5 AK)
fglasqm	Glasfläche (<=2000 qm, >2000 - <=5000 qm, >5000 qm)
fjahr	Erhebungsjahr (1992, 1993, 1994)
funtert	Unternehmensertrag (<=200000 DM, >200000 - <=500000 DM, >500000 DM)
given.ak.3	shingle Anzahl Arbeitskräfte (Intervalle siehe Übersicht B8)
given.ak.4	shingle Anzahl Arbeitskräfte (Intervalle siehe Übersicht B8)
given.glasqm.3	shingle qm Glasfläche (Intervalle siehe Übersicht B8)
given.glasqm.4	shingle qm Glasfläche (Intervalle siehe Übersicht B8)
given.untert.3	shingle Unternehmensertrag (Intervalle siehe Übersicht B8)
given.untert.4	shingle Unternehmensertrag (Intervalle siehe Übersicht B8)
region	Region (übrige Regionen, Region 1)

Strukturdaten

ak	Arbeitskräfte insgesamt, Kennzahl 20
epertp	Erträge aus Eigenproduktion in % BE, Kennzahl 32[11)]
eqm	Einheitsquadratmeter, Kennzahl 3
fremdakp	Fremd-AK in % der AK, Kennzahl 21
glasqm	Glasfläche in qm, Kennzahl 2
glasqmak	Glasfläche je AK in qm, Kennzahl 23

Investitions- und Vermögensdaten

anvermp	Anlagevermögen (ohne Boden) in % des Vermögens, Kennzahl 10
fkp	Fremdkapital in % Anlagevermögen (ohne Boden), Kennzahl 14
netinvp	Nettoinvestitionen in % Anlagevermögen (ohne Boden), Kennzahl 27
verm	Vermögen in 1000 DM, Kennzahl 8

Aufwandsdaten

allgawp	allgemeiner Aufwand in % BE, Kennzahl 59
heizp	Heizmaterial in % BE, Kennzahl 45
heizqm	Heizmaterial je qm heizbare Glasfläche (DM), Kennzahl 61
lohnak	Lohn je entlohnte AK (DM), Kennzahl 60
lohnqp	Lohnquote in % BE, Kennzahl 50
saatp	Saat- und Pflanzgut in % BE, Kennzahl 43
spezp	Spezialaufwand Eigenproduktion in % BE, Kennzahl 47

Erfolgsdaten[2]

beinkak	Betriebseinkommen je AK (DM), Kennzahl 75
beinkeqm	Betriebseinkommen je Eqm (DM),Kennzahl 78
beinkp	Betriebseinkommen in % BE, Kennzahl 69
kapkoef	Kapitalkoeffizient, Kennzahl 81
rdiffp	Reinertragsdifferenz in % BE, Kennzahl 81
rentkoef	Rentabilitätskoeffizient, Kennzahl 82
rtak	Reinertrag je AK (DM), Kennzahl 76
rteqm	Reinertrag je Eqm (DM), Kennzahl 79
rtp	Reinertrag in % BE, Kennzahl, 71

[1] BE steht für Betriebsertrag

[2] Erfolgsdaten im weiteren und umgangssprachlichen Sinne; sprachlich korrekter wäre die Unterteilung in Produktivitäts- und Rentabilitätskennzahlen, wobei letztere Erfolgskennzahlen im engeren Sinne darstellen.

Übersicht B2: Gruppierungsdaten (Anzahl Fälle in den Kategorien)

	Nobservd
abswg	
1.00	710
2.00	101
3.00	80

	Nobservd
fak	
lt 2	69
ge 2 to lt 5	402
ge 5	420

	Nobservd
fglasqm	
lt 2000	80
ge 2000 to lt 5000	378
ge 5000	433

	Nobservd
fjahr	
92.00	297
93.00	297
94.00	297

	Nobservd
fregion	
uebrige Regionen	420
Region 1	471

	Nobservd
funtert	
lt 200000	61
ge 200000 to lt 500000	283
ge 500000	547

Übersicht B3: Univariate Statistiken der Strukturdaten

Summary statistics for ak	Summary statistics for epertp	Summary statistics for eqm
Mean = 5.331	Mean = 92.717	Mean = 140403.336
Median = 4.630	Median = 96.840	Median = 118000.000
Minimum = 1.000	Minimum = 61.540	Minimum = 11000.000
Maximum = 18.220	Maximum = 100.000	Maximum = 528200.000
Range = 17.220	Range = 38.460	Range = 517200.000
Lower quartile = 3.000	Lower quartile = 90.240	Lower quartile = 75250.000
Upper quartile = 6.840	Upper quartile = 98.665	Upper quartile = 184728.000
Standard deviation = 3.082	Standard deviation = 8.877	Standard deviation = 89523.951
Skewness = 1.187	Skewness = -1.613	Skewness = 1.257
Kurtosis = 1.407	Kurtosis = 1.744	Kurtosis = 1.549

Summary statistics for fremdakp	Summary statistics for glasqm	Summary statistics for glasqmak
Mean = 61.770	Mean = 5769.003	Mean = 1214.876
Median = 66.670	Median = 4900.000	Median = 1128.435
Minimum = 0.000	Minimum = 450.000	Minimum = 56.729
Maximum = 100.000	Maximum = 20000.000	Maximum = 6655.556
Range = 100.000	Range = 19550.000	Range = 6598.826
Lower quartile = 49.580	Lower quartile = 3350.000	Lower quartile = 745.455
Upper quartile = 78.570	Upper quartile = 7500.000	Upper quartile = 1538.462
Standard deviation = 21.899	Standard deviation = 3520.436	Standard deviation = 654.557
Skewness = -0.924	Skewness = 1.236	Skewness = 1.730
Kurtosis = 0.381	Kurtosis = 1.691	Kurtosis = 7.660

Für alle Variablen gilt: Number of observations = 891 Number of missing values = 0
Standard Error of Skewness = 0.082 Standard Error of Kurtosis = 0.164

Übersicht B4: Univariate Statistiken der Investitions- und Vermögensdaten

Summary statistics for anvermp	Summary statistics for fkp	Summary statistics for netinvp
Mean = 44.929	Mean = 161.169	Mean = -0.865
Median = 45.460	Median = 109.630	Median = -5.810
Minimum = 0.000	Minimum = 0.000	Minimum = -318.010
Maximum = 93.610	Maximum = 3019.670	Maximum = 241.640
Range = 93.610	Range = 3019.670	Range = 559.650
Lower quartile = 36.800	Lower quartile = 58.460	Lower quartile = -14.800
Upper quartile = 54.430	Upper quartile = 186.580	Upper quartile = 9.530
Standard deviation = 13.839	Standard deviation = 217.112	Standard deviation = 30.840
Skewness = -0.366	Skewness = 6.121	Skewness = 0.459
Kurtosis = 0.420	Kurtosis = 56.958	Kurtosis = 23.771
Summary statistics for verm		
Mean = 721.213		
Median = 596.000		
Minimum = 67.000		
Maximum = 3811.000		
Range = 3744.000		
Lower quartile = 383.000		
Upper quartile = 922.000		
Standard deviation = 486.359		
Skewness = 1.769		
Kurtosis = 4.841		

Für alle Variablen gilt: Number of observations = 891 Number of missing values = 0
Standard Error of Skewness = 0.082 Standard Error of Kurtosis = 0.164

Übersicht B5: Univariate Statistiken der Aufwandsdaten

Summary statistics for allgawp

Mean	26.576
Median	25.700
Minimum	7.130
Maximum	64.480
Range	57.350
Lower quartile	20.730
Upper quartile	31.950
Standard deviation	8.208
Skewness	0.611
Kurtosis	0.715

Summary statistics for heizp

Mean	7.570
Median	6.860
Minimum	0.000
Maximum	33.140
Range	33.140
Lower quartile	4.740
Upper quartile	9.555
Standard deviation	4.141
Skewness	1.561
Kurtosis	5.184

Summary statistics for heizqm

Mean	9.511
Median	8.270
Minimum	0.000
Maximum	37.350
Range	37.350
Lower quartile	5.390
Upper quartile	12.220
Standard deviation	5.644
Skewness	1.177
Kurtosis	1.542

Summary statistics for lohnak

Mean	33134.164
Median	32908.610
Minimum	0.000
Maximum	114511.000
Range	114511.000
Lower quartile	25192.510
Upper quartile	40267.630
Standard deviation	12666.877
Skewness	0.583
Kurtosis	3.578

Summary statistics for lohnqp

Mean	35.074
Median	32.350
Minimum	11.250
Maximum	242.720
Range	231.470
Lower quartile	26.560
Upper quartile	40.270
Standard deviation	14.591
Skewness	4.906
Kurtosis	54.190

Summary statistics for saatp

Mean	11.747
Median	9.770
Minimum	0.000
Maximum	159.300
Range	159.300
Lower quartile	4.090
Upper quartile	16.950
Standard deviation	10.849
Skewness	3.734
Kurtosis	38.996

Summary statistics for spezp

Mean	33.036
Median	32.290
Minimum	6.080
Maximum	248.160
Range	242.080
Lower quartile	24.700
Upper quartile	39.640
Standard deviation	12.996
Skewness	5.399
Kurtosis	83.239

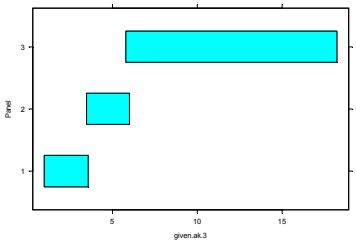
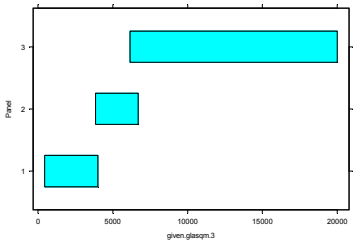
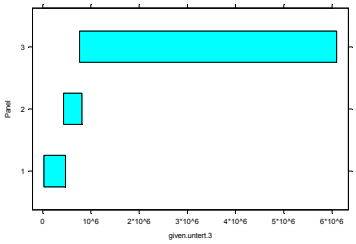
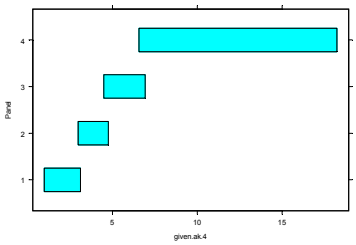
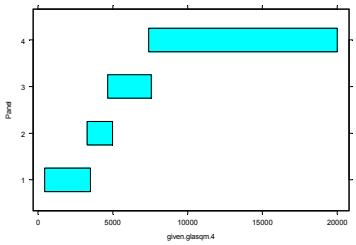
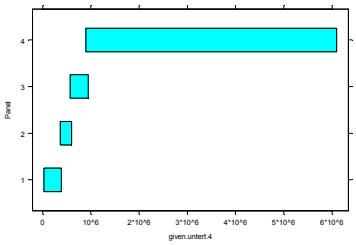
Für alle Variablen gilt: Number of observations = 891 Number of missing values = 0
Standard Error of Skewness = 0.082 Standard Error of Kurtosis = 0.164

Übersicht B6: Univariate Statistiken der Erfolgsdaten

Summary statistics for beinkak Mean = 49215.359 Median = 46764.426 Minimum = -206170.233 Maximum = 162085.671 Range = 368255.904 Lower quartile = 35091.367 Upper quartile = 60218.297 Standard deviation = 23791.709 Skewness = -0.563 Kurtosis = 15.434	Summary statistics for beinkeqm Mean = 2.218 Median = 1.721 Minimum = -1.367 Maximum = 29.878 Range = 31.244 Lower quartile = 1.228 Upper quartile = 2.551 Standard deviation = 2.098 Skewness = 6.258 Kurtosis = 65.276	Summary statistics for beinkp Mean = 37.570 Median = 37.710 Minimum = -206.390 Maximum = 72.750 Range = 279.140 Lower quartile = 31.500 Upper quartile = 44.890 Standard deviation = 13.768 Skewness = -6.472 Kurtosis = 110.045
Summary statistics for kapkoef Mean = 3.247 Median = 2.770 Minimum = -94.300 Maximum = 49.980 Range = 144.280 Lower quartile = 2.130 Upper quartile = 3.635 Standard deviation = 4.547 Skewness = -8.954 Kurtosis = 255.453	Summary statistics for rdiffp Mean = -4.100 Median = -1.720 Minimum = -272.580 Maximum = 42.560 Range = 315.140 Lower quartile = -9.905 Upper quartile = 6.400 Standard deviation = 21.441 Skewness = -5.159 Kurtosis = 56.038	Summary statistics for rentkoef Mean = 0.982 Median = 0.960 Minimum = -3.120 Maximum = 2.980 Range = 6.100 Lower quartile = 0.770 Upper quartile = 1.200 Standard deviation = 0.403 Skewness = -0.625 Kurtosis = 13.224
Summary statistics for rtak Mean = 7771.993 Median = 5666.621 Minimum = -256326.769 Maximum = 110977.995 Range = 367304.764 Lower quartile = -3565.326 Upper quartile = 17742.496 Standard deviation = 21958.218 Skewness = -1.261 Kurtosis = 24.735	Summary statistics for rteqm Mean = 0.257 Median = 0.207 Minimum = -4.337 Maximum = 14.058 Range = 18.395 Lower quartile = -0.163 Upper quartile = 0.614 Standard deviation = 0.995 Skewness = 3.926 Kurtosis = 49.478	Summary statistics for rtp Mean = 2.495 Median = 4.640 Minimum = -256.600 Maximum = 47.470 Range = 304.070 Lower quartile = -3.220 Upper quartile = 12.400 Standard deviation = 20.474 Skewness = -5.305 Kurtosis = 58.422

Für alle Variablen gilt: Number of observations = 891 Number of missing values = 0
Standard Error of Skewness = 0.082 Standard Error of Kurtosis = 0.164

Übersicht B7: Einteilung von Glasfläche, Anzahl Arbeitskräfte und Unternehmensertrag in drei beziehungsweise vier Gruppen mit Hilfe des equal-count-Algorithmus bei einer angestrebten Überlappung von 10%

Arbeitskräfte	Glasfläche	Unternehmensertrag
Intervals: min max count 1.00 3.57 318 3.48 6.00 319 5.80 18.22 318 Overlap between adjacent intervals: [1] 32 32	Intervals: min max count 450 4000 335 3839 6700 322 6150 20000 318 Overlap between adjacent intervals: [1] 49 35	Intervals: min max count 29003 469423 318 431146 810450 319 766122 6096290 318 Overlap between adjacent intervals: [1] 32 32
		
Intervals: min max count 1.00 3.13 244 3.00 4.77 244 4.51 6.93 243 6.58 18.22 241 Overlap between adjacent intervals: [1] 30 26 25	Intervals: min max count 450 3500 256 3300 4980 252 4665 7580 243 7400 20000 242 Overlap between adjacent intervals: [1] 50 26 26	Intervals: min max count 29003 384083 241 364982 596013 241 570989 941558 241 899957 6096290 241 Overlap between adjacent intervals: [1] 24 25 24
		

Übersicht B8: Tests auf univariate (marginal) und multivariate (radius) Normalverteilung

	Type of test	Variate(s)	Test statistic: Anderson-Darling	Test statistic: Cramer-von Mises	Test statistic: Watson
ak	Marginal	1	20.173 **	3.129 **	2.356 **
allgawp	Marginal	2	3.187 **	0.495 **	0.346 **
anvermp	Marginal	3	1.914 **	0.254 **	0.188 **
beinkak	Marginal	4	10.662 **	1.758 **	1.585 **
beinkeqm	Marginal	5	77.241 **	14.330 **	13.176 **
beinkp	Marginal	6	15.806 **	2.409 **	2.322 **
epertp	Marginal	7	76.314 **	13.733 **	11.811 **
eqm	Marginal	8	23.852 **	3.864 **	2.970 **
fkp	Marginal	9	94.475 **	17.309 **	15.957 **
fremdakp	Marginal	10	16.625 **	2.468 **	1.870 **
glasqm	Marginal	11	21.582 **	3.754 **	2.967 **
glasqmak	Marginal	12	11.477 **	1.786 **	1.330 **
heizqm	Marginal	13	20.557 **	3.453 **	2.688 **
heizp	Marginal	14	13.083 **	2.081 **	1.548 **
kapkoef	Marginal	15	144.554 **	27.974 **	27.913 **
lohnak	Marginal	16	3.335 **	0.517 **	0.503 **
lohnqp	Marginal	17	30.200 **	4.976 **	4.343 **
netinvp	Marginal	18	39.353 **	7.045 **	6.855 **
rdiffp	Marginal	19	34.665 **	6.098 **	5.616 **
rentkoef	Marginal	20	7.178 **	1.258 **	1.215 **
rtak	Marginal	21	13.349 **	2.281 **	2.197 **
rteqm	Marginal	22	28.060 **	4.866 **	4.759 **
rtp	Marginal	23	35.730 **	6.227 **	5.768 **
saatp	Marginal	24	21.881 **	3.148 **	2.538 **
spezp	Marginal	25	11.446 **	1.526 **	1.350 **
verm	Marginal	26	28.585 **	4.755 **	3.738 **

Type of test	Variate(s)	Test statistic: Anderson-Darling	Test statistic: Cramer-von Mises	Test statistic: Watson
Radius		720.130 **	80.653 **	23.303 **

?, *, ** indicate significance at 10%, 5% and 1% levels respectively

Übersicht B9: Univariate Statistiken einiger ausgewählter Kennzahlen vor und nach dem Ausschluß extremer Werte

vor Ausschluß extremer Werte

identifier	missing	missing cases	length	mean	std.dev.	median	min.	max.
allgawp	0		891.0	26.58	8.208	25.70	7.130	64.48
anvermp	0		891.0	44.93	13.84	45.46	0	93.61
beinkak	0		891.0	49215	23971.7	46764	-206170	162086
beinkeqm	0		891.0	2.218	2.098	1.721	-1.367	29.88
beinkp	0		891.0	37.57	13.77	37.71	-206.4	72.75
glasqmak	0		891.0	1215	654.6	1128	56.73	6656
heizqm	0		891.0	9.511	5.644	8.270	0	37.35
lohnak	0		891.0	33134	12666.9	32909	0	114511
lohnqp	0		891.0	35.07	14.59	32.35	11.25	242.7
rentkoef	0		891.0	0.9825	0.403	0.9600	-3.120	2.980
rdiffp	0		891.0	-4.100	21.44	-1.720	-272.6	42.56
spezp	0		891.0	33.04	12.99	32.29	6.080	248.2

nach Ausschluß extremer Werte

identifier	missing	missing cases	length	mean	std.dev.	median	min.	max.
allgawp	0		891.0	26.58	8.208	25.70	7.130	64.48
anvermp	0		891.0	44.93	13.84	45.46	0	93.61
beinkak	1	557	891.0	49502	7084.1	46852	-28691	162086
beinkeqm	3	811,812,813	891.0	2.138	1.562	1.719	-1.367	11.52
beinkp	1	557	891.0	37.84	11.08	37.72	-31.44	72.75
glasqmak	4	349,401,557,558	891.0	1197	598.7	1128	56.73	3929
heizqm	0		891.0	9.511	5.644	8.270	0	37.35
lohnak	3	316,361,362	891.0	32895	11988.8	32868	0	78459
lohnqp	3	484,485,486	891.0	34.60	11.74	32.30	11.25	102.6
rentkoef	1	557	891.0	0.9871	0.379	0.9600	-0.5500	2.980
rdiffp	3	485,486,557	891.0	-3.347	16.70	-1.690	-91.94	42.56
spezp	1	557	891.0	32.79	10.82	32.27	6.080	84.20

Übersicht B10: Rangkorrelationen der Kennzahlen der vier Variablensets untereinander (Korrelationen von mehr als 0,7 sind fett geschrieben)

Spearman Rank Correlation

*** Correlation matrix (adjusted for ties) ***
Sample size: 891

Korrelationen der Erfolgsdaten untereinander

1	1.000								
2	0.341	1.000							
3	0.536	0.408	1.000						
4	-0.503	-0.451	**-0.742**	1.000					
5	**0.846**	0.381	0.627	-0.664	1.000				
6	**0.853**	0.391	0.652	-0.680	**0.994**	1.000			
7	**0.905**	0.344	0.564	-0.551	**0.971**	**0.974**	1.000		
8	**0.814**	0.487	0.573	-0.587	**0.943**	**0.941**	**0.935**	1.000	
9	**0.859**	0.356	0.629	-0.596	**0.990**	**0.985**	**0.982**	**0.944**	1.000

1	2	3	4	5	6	7	8	9

1 beinkak 2 beinkeqm 3 beinkp 4 kapkoef 5 rdiffp 6 rentkoef 7 rtak 8 rteqm 9 rtp

Korrelationen der Strukturdaten untereinander

10	1.000					
11	0.005	1.000				
12	0.540	0.384	1.000			
13	**0.807**	-0.028	0.423	1.000		
14	0.603	0.346	**0.941**	0.474	1.000	
15	-0.336	0.363	0.497	-0.286	0.491	1.000

10	11	12	13	14	15

10 ak 11 epertp 12 eqm 13 fremdakp 14 glasqm 15 glasqmak

Korrelationen der Investitions- und Vermögensdaten untereinander

16	1.000			
17	-0.296	1.000		
18	0.237	-0.115	1.000	
19	0.426	-0,.80	0.199	1.000

16	17	18	19

16 anvermp 17 fkp 18 netinvp 19 verm

Korrelationen der Aufwandsdaten untereinander

20	1.000						
21	-0.023	1.000					
22	-0.265	0.587	1.000				
23	-0.126	-0.012	0.084	1.000			
24	-0.029	0.258	0.110	0.150	1.000		
25	-0.293	0.080	0.280	-0.036	-0.264	1.000	
26	-0.180	0.401	0.252	-0.086	-0.168	**0.737**	1.000

20	21	22	23	24	25	26

20 allgawp 21 heizp 22 heizqm 23 lohnak 24 lohnqp 25 saatp 26 spezp

Übersicht B11: Rangkorrelationen zwischen den Erfolgskennzahlen und den Kennzahlen der übrigen Datensets (Korrelationen von mehr als 0,7 sind fett geschrieben)

Spearman Rank Correlation

*** Correlation matrix (adjusted for ties) ***
Sample size: 891

Korrelationen der Erfolgsdaten und der Strukturdaten

10	0.004	0.360	0.128	-0.174	0.134	0.134	0.083	0.188	0.119
11	0.205	-0.247	0.072	-0.060	0.206	0.207	0.219	0.138	0.215
12	0.332	-0.278	0.097	-0.086	0.304	0.304	0.331	0.228	0.318
13	0.032	0.333	0.123	-0.146	0.165	0.166	0.125	0.218	0.156
14	0.299	-0.182	0.022	-0.062	0.256	0.253	0.284	0.206	0.263
15	0.390	-0.604	-0.066	0.065	0.225	0.220	0.303	0.116	0.249

1	2	3	4	5	6	7	8	9
10 ak	11 epertp	12 eqm	13 fremdakp	14 glasqm	15 glasqmak			

Korelation der Erfolgsdaten mit den Investitions- und Vermögensdaten

16	0.174	0.088	0.002	0.410	0.055	0.054	0.166	0.137	0.150
17	-0.249	-0.192	-0.212	0.064	-0.220	-0.220	-0.242	-0.239	-0.248
18	0.148	-0.014	0.014	0.063	0.144	0.138	0.175	0.156	0.167
19	0.272	0.234	-0.048	0.140	0.190	0.182	0.249	0.255	0.228

1	2	3	4	5	6	7	8	9
16 anvermp	17 fkp	18 netinvp	19 verm					

Korrelationen der Erfolgsdaten und der Aufwandsdaten

20	-0.226	-0.473	-0.397	0.442	-0.289	-0.287	-0.226	-0.292	-0.256
21	-0.339	-0.282	-0.264	0.223	-0.382	-0.377	-0.383	-0.376	-0.384
22	-0.234	0.424	-0.190	0.087	-0.238	-0.242	-0.263	-0.135	-0.259
23	0.359	0.236	0.174	-0.131	0.019	0.031	0.055	0.054	0.019
24	-0.564	-0.113	0.144	0.026	-0.611	-0.584	-0.670	-0.597	-0.621
25	-0.126	0.110	-0.499	0.236	-0.163	-0.184	-0.144	-0.110	-0.185
26	-0.282	-0.277	-0.660	0.404	-0.345	-0.365	-0.313	-0.318	-0.362

1	2	3	4	5	6	7	8	9
20 allgawp	21 heizp	22 heizqm	23 lohnak	24 lohnqp	25 saatp	26 spezp		

Korrelationen einiger Aufwandsdaten und der Strukturdaten

20	-0.159	-0.020	0.121	-0.157	0.067	0.270	1.000		
24	-0.029	-0.181	-0.300	-0.069	-0.316	-0.364	-0.029	1.000	
26	-0.090	0.268	0.001	-0.079	0.077	0.109	-0.180	-0.168	1.000

10	11	12	13	14	15	20	24	26
20 allgawp	24 lohnqp	26 spezp						
10 ak	11 epertp	12 eqm	13 fremdakp	14 glasqm	15 glasqmak			

Korrelationen einiger Aufwandsdaten und der Vermögensdaten

20	0.180	-0.101	0.057	0.010	1.000		
24	-0.220	0.114	-0.216	-0.373	-0.029	1.000	
26	-0.151	0.297	-0.041	0.017	-0.180	-0.168	1.000

16	17	18	19	20	24	26
20 allgawp	24 lohnqp	26 spezp				
16 anvermp	17 fkp	18 netinvp	19 verm			

Übersicht B12: Gruppenbildung der Kennzahlenbetriebe nach Glasfläche, Region und Erhebungsjahr (Anzahl Fälle je Gruppe)

Glasfläche (shingle)

Gruppe	1	2	3	4	Nobservd
1	57	0	0	0	57
2	51	0	0	0	51
3	50	0	0	0	50
4	34	0	0	0	34
5	33	0	0	0	33
6	31	0	0	0	31
7	0	30	0	0	30
8	0	32	0	0	32
9	0	33	0	0	33
10	0	38	0	0	38
11	0	35	0	0	35
12	0	34	0	0	34
13	0	0	24	0	24
14	0	0	24	0	24
15	0	0	24	0	24
16	0	0	48	0	48
17	0	0	49	0	49
18	0	0	48	0	48
19	0	0	0	29	29
20	0	0	0	33	33
21	0	0	0	33	33
22	0	0	0	37	37
23	0	0	0	40	40
24	0	0	0	44	44
	256	202	217	216	891

Region

Gruppe	uebrige Laender	Region 1	Nobservd
1	57	0	57
2	51	0	51
3	50	0	50
4	0	34	34
5	0	33	33
6	0	31	31
7	30	0	30
8	32	0	32
9	33	0	33
10	0	38	38
11	0	35	35
12	0	34	34
13	24	0	24
14	24	0	24
15	24	0	24
16	0	48	48
17	0	49	49
18	0	48	48
19	29	0	29
20	33	0	33
21	33	0	33
22	0	37	37
23	0	40	40
24	0	44	44
	420	471	891

Jahr

Gruppe	1992	1993	1994	Nobservd
1	57	0	0	57
2	0	51	0	51
3	0	0	50	50
4	34	0	0	34
5	0	33	0	33
6	0	0	31	31
7	30	0	0	30
8	0	32	0	32
9	0	0	33	33
10	38	0	0	38
11	0	35	0	35
12	0	0	34	34
13	24	0	0	24
14	0	24	0	24
15	0	0	24	24
16	48	0	0	48
17	0	49	0	49
18	0	0	48	48
19	29	0	0	29
20	0	33	0	33
21	0	0	33	33
22	37	0	0	37
23	0	40	0	40
24	0	0	44	44
	297	297	297	891

Zusammenfassung

Gruppe	Fläche	Region	Jahr	Anzahl
1	1	übrige	1992	57
2	1	übrige	1993	51
3	1	übrige	1994	50
4	1	Region 1	1992	34
5	1	Region 1	1993	33
6	1	Region 1	1994	31
7	2	übrige	1992	30
8	2	übrige	1993	32
9	2	übrige	1994	33
10	2	Region 1	1992	38
11	2	Region 1	1993	35
12	2	Region 1	1994	34
13	3	übrige	1992	24
14	3	übrige	1993	24
15	3	übrige	1994	24
16	3	Region 1	1992	48
17	3	Region 1	1993	49
18	3	Region 1	1994	48
19	4	übrige	1992	29
20	4	übrige	1993	33
21	4	übrige	1994	33
22	4	Region 1	1992	37
23	4	Region 1	1993	40
24	4	Region 1	1994	44

Übersicht B13: Schätzung der Parameter (k = Form-, b = Skalenparameter) der Gamma-Verteilung aus den Abweichungen der Eigenvektoren der 24 Gruppen der Kennzahlenbetriebe vom ‘typischen’ Eigenvektor

**** first eigenvector ****

***** Fit continuous distribution *****

*** Sample Statistics ***

Sample Size	24		
Mean	0.16	Variance	0.02
Skewness	1.85	Kurtosis	2.66
Quartiles:	25%	50%	75%
	0.1	0.1	0.2

*** Summary of analysis ***
Observations: dist
Parameter estimates from individual data values
Distribution: Gamma
f(x) = (b**k).(x**(k-1)).exp(-bx)/Gamma(k), x>0
Deviance: 0.86 on 2 d.f.

*** Estimates of parameters ***

	estimate	s.e.	Correlations	
k	**1.5868**	**0.4191**	**1.0000**	
b	**10.1209**	**3.1366**	**0.8523**	**1.0000**

*** Fitted quartiles ***

25%	50%	75%
0.066	0.125	0.214

*** Fitted values (expected frequencies) and residuals ***

x	Number Observed	Number Expected	Weighted Residual
< 0.047	4	3.91	0.05
< 0.083	5	3.93	0.52
< 0.128	5	4.42	0.27
< 0.199	5	4.96	0.02
> 0.199	5	6.78	-0.72

**** second eigenvector ****

***** Fit continuous distribution *****

*** Sample Statistics ***

Sample Size	24		
Mean	0.58	Variance	0.14
Skewness	0.79	Kurtosis	-0.07
Quartiles:	25%	50%	75%
	0.3	0.5	0.9

*** Summary of analysis ***
Observations: dist
Parameter estimates from individual data values
Distribution: Gamma
f(x) = (b**k).(x**(k-1)).exp(-bx)/Gamma(k), x>0
Deviance: 0.77 on 2 d.f.

*** Estimates of parameters ***

	estimate	s.e.	Correlations	
k	**2.3266**	**0.6297**	**1.0000**	
b	**3.9975**	**1.2071**	**0.8964**	**1.0000**

*** Fitted quartiles ***

25%	50%	75%
0.301	0.501	0.775

*** Fitted values (expected frequencies) and residuals ***

x	Number Observed	Number Expected	Weighted Residual
< 0.23	4	3.93	0.03
< 0.44	5	6.43	-0.59
< 0.60	5	4.12	0.42
< 0.92	5	5.47	-0.20
> 0.92	5	4.05	0.46

**** third eigenvector ****

***** Fit continuous distribution *****

*** Sample Statistics ***

Sample Size	24		
Mean	0.82	Variance	0.23
Skewness	0.37	Kurtosis	-0.81
Quartiles:	25%	50%	75%
	0.4	0.8	1.2

*** Summary of analysis ***
Observations: dist
Parameter estimates from individual data values
Distribution: Gamma
f(x) = (b**k).(x**(k-1)).exp(-bx)/Gamma(k), x>0
Deviance: 3.91 on 2 d.f.

*** Estimates of parameters ***

	estimate	s.e.	Correlations	
k	**2.5196**	**0.6850**	**1.0000**	
b	**3.0845**	**0.9277**	**0.9039**	**1.0000**

*** Fitted quartiles ***

25%	50%	75%
0.438	0.712	1.082

*** Fitted values (expected frequencies) and residuals ***

x	Number Observed	Number Expected	Weighted Residual
< 0.35	4	4.08	-0.04
< 0.57	5	4.82	0.08
< 1.03	5	8.42	-1.28
< 1.24	5	2.39	1.47
> 1.24	5	4.29	0.33

**** fourth eigenvector ****

***** Fit continuous distribution *****

*** Sample Statistics ***

Sample Size	24		
Mean	1.04	Variance	0.20
Skewness	0.35	Kurtosis	-1.28
Quartiles:	25%	50%	75%
	0.7	0.9	1.5

*** Summary of analysis ***
Observations: dist
Parameter estimates from individual data values
Distribution: Gamma
f(x) = (b**k).(x**(k-1)).exp(-bx)/Gamma(k), x>0
Deviance: 3.54 on 2 d.f.

*** Estimates of parameters ***

	estimate	s.e.	Correlations	
k	**5.3403**	**1.4967**	**1.0000**	
b	**5.1510**	**1.5138**	**0.9537**	**1.0000**

*** Fitted quartiles ***

25%	50%	75%
0.710	0.973	1.294

*** Fitted values (expected frequencies) and residuals ***

x	Number Observed	Number Expected	Weighted Residual
< 0.61	4	3.93	0.04
< 0.81	5	4.25	0.36
< 1.00	5	4.41	0.27
< 1.59	5	8.68	-1.36
> 1.59	5	2.74	1.22

Übersicht B14: Ergebnisse des Gruppenanalysemodels der 24 Gruppen der Kennzahlenbetriebe; mittlere Koeffizienten der ersten vier Hauptkomponenten

*** Between-Groups Comparison of Principal Components ***

average component loadings b that minimize V
directions closest to each subspace

varname	b[1]	b[2]	b[3]	b[4]
allgawp	0.2357	-0.3868	-0.0770	-0.2056
spezp	0.2180	0.1731	-0.1986	0.5007
lohnqp	0.1874	0.2726	0.3642	-0.4040
lohnak	-0.0271	-0.0201	0.0290	-0.2747
heizqm	-0.0102	0.3571	-0.3305	0.1157
eqm	-0.0242	-0.2621	0.4273	0.1097
glasqm	-0.0186	-0.2195	0.3220	0.2151
glasqmak	0.0653	-0.3908	0.0683	0.2796
fkp	0.1187	0.2351	0.2858	0.3352
anvermp	-0.0005	-0.3573	-0.3363	-0.2536
beinkp	-0.3654	0.0734	0.2928	-0.2471
beinkak	-0.3541	-0.1997	-0.0917	0.0843
beinkeqm	-0.3449	0.2209	-0.2575	-0.1443
kapkoef	0.3379	-0.2118	-0.2472	-0.0722
rentkoef	-0.4207	-0.1084	-0.0284	0.1441
rdiffp	-0.4154	-0.1220	-0.0618	0.1636

Variablen sortiert nach erster mittlerer Hauptkomponente (Erfolgs- und Kostendimension)

varname	b[1]
rentkoef	-0.4207
rdiffp	-0.4154
beinkp	-0.3654
beinkak	-0.3541
beinkeqm	-0.3449
kapkoef	0.3379
lohnak	-0.0271
eqm	-0.0242
glasqm	-0.0186
heizqm	-0.0102
anvermp	-0.0005
glasqmak	0.0653
fkp	0.1187
lohnqp	0.1874
spezp	0.2180
allgawp	0.2357

Variablen sortiert nach zweiter mittlerer Hauptkomponente (Technologie- und Vermögensdimension)

varname	b[2]
glasqmak	-0.3908
allgawp	-0.3868
anvermp	-0.3573
eqm	-0.2621
glasqm	-0.2195
beinkak	-0.1997
rdiffp	-0.1220
rentkoef	-0.1084
lohnak	-0.0201
beinkp	0.0734
spezp	0.1731
kapkoef	-0.2118
beinkeqm	0.2209
fkp	0.2351
lohnqp	0.2726
heizqm	0.3571

Übersicht B15: Ergebnisse des Gruppenanalysemodells der 24 Gruppen der Kennzahlenbetriebe; Winkel (delta) jeder Gruppe zur mittleren Konfiguration in den ersten vier Dimensionen

*** Between-Groups Comparison of Principal Components ***

angles formed by each group with each direction

group	delta[1]	delta[2]	delta[3]	delta[4]
1	10.23	16.05	50.86	12.87
2	12.33	17.15	30.17	14.23
3	15.89	18.25	22.55	15.41
4	8.51	18.33	15.46	38.59
5	14.92	25.60	40.86	41.18
6	37.68	34.20	41.67	59.87
7	7.08	12.35	31.44	49.52
8	11.09	11.78	60.39	60.70
9	8.36	12.74	74.14	27.87
10	12.65	25.17	22.59	26.71
11	11.12	25.16	18.75	59.48
12	8.55	15.57	48.71	43.96
13	21.61	66.84	28.69	36.33
14	11.43	50.08	26.43	17.55
15	8.30	21.29	34.56	16.07
16	14.64	18.93	29.83	32.65
17	12.37	17.02	18.88	54.25
18	15.21	20.17	38.59	48.51
19	8.37	21.95	37.78	25.33
20	10.20	15.02	18.05	28.56
21	16.63	17.91	8.20	36.96
22	9.28	15.59	18.40	49.89
23	9.58	18.73	25.78	38.05
24	16.86	21.15	23.36	35.68

sum of squared cosines				
sscos	22.76	19.97	16.86	15.15

the groups are defined by	qm1_shn	fregion	fjahr
with levels	4.000	2.000	3.000

Übersicht B16a: Vergleich der Hauptkomponentenanalyseergebnisse für Gruppe 7 und Gruppe 6 der 24 Gruppen der Kennzahlenbetriebe

```
*****      Analysis of Subgroup           7       *****

Total number of units in data =  30

No outliers were detected
Number of iterations used =  1

 No     Root    %%  Cum   %  Scree Diagram (* represents 1%)
  1   5.0062  313  313  31 ******************************
  2   3.4947  218  531  22 **********************
  3   1.8885  118  649  12 ************
  4   1.5689   98  747  10 **********
  5   1.0490   66  813   7 *******
  6   0.8460   53  866   5 *****
  7   0.6660   42  907   4 ****
  8   0.5913   37  944   4 ****
  9   0.4514   28  973   3 ***
 10   0.2161   14  986   1 *
 11   0.1017    6  992   1 *
 12   0.0598    4  996   0
 13   0.0346    2  998   0
 14   0.0219    1 1000   0
 15   0.0034    0 1000   0
 16   0.0005    0 1000   0

Scale:  1 asterisk represents 1 unit.


l[7]['Vectors']
               1           2           3           4

  1      0.1017      0.4256     -0.2232      0.0200
  2      0.1894     -0.0899      0.5067      0.0140
  3      0.2476     -0.2377     -0.3031     -0.1737
  4     -0.0733     -0.1649      0.4446      0.2986
  5      0.0872     -0.3448      0.1474      0.1450
  6     -0.0534      0.1482      0.2422     -0.5392
  7     -0.0846      0.0531      0.3169     -0.5099
  8     -0.0357      0.3741      0.2220     -0.0691
  9      0.1241     -0.2543     -0.1641     -0.3347
 10     -0.1227      0.4383     -0.0964      0.2073
 11     -0.3760     -0.0641     -0.2856     -0.1220
 12     -0.3899      0.0408      0.2185      0.1123
 13     -0.2815     -0.2672     -0.0498      0.2604
 14      0.2955      0.3295      0.0553      0.2050
 15     -0.4377      0.0196     -0.0071     -0.0124
 16     -0.4316      0.0244     -0.0513     -0.1050

 1     allgawp   2       spezp   3      lohnqp
 4      lohnak   5      heizqm   6         eqm
 7      glasqm   8    glasqmak   9         fkp
10     anvermp  11      beinkp  12     beinkak
13    beinkeqm  14     kapkoef  15      rdiffp
16    rentkoef


The first two principal components explain       53.13
percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

     allgawp      0.1915
       spezp      0.0439
      lohnqp      0.1178
      lohnak      0.0326
      heizqm      0.1265
         eqm      0.0248
      glasqm      0.0100
    glasqmak      0.1412
         fkp      0.0801
     anvermp      0.2072
      beinkp      0.1455
     beinkak      0.1537
    beinkeqm      0.1506
     kapkoef      0.1959
      rdiffp      0.1919
    rentkoef      0.1868
```

```
*****      Analysis of Subgroup           6       *****

Total number of units in data =  31

 No outliers were detected
 Number of iterations used =  1

No     Root    %%  Cum   %  Scree Diagram (* represents 2%)
   1   6.0999  381  381  38 ******************
   2   2.1535  135  516  13 *******
   3   1.7955  112  628  11 ******
   4   1.2453   78  706   8 ****
   5   1.1292   71  776   7 ****
   6   1.0141   63  840   6 ***
   7   0.7906   49  889   5 ***
   8   0.6141   38  928   4 **
   9   0.5248   33  960   3 **
  10   0.2635   16  977   2 *
  11   0.2130   13  990   1 *
  12   0.0810    5  995   1 *
  13   0.0383    2  998   0
  14   0.0290    2  999   0
  15   0.0081    1 1000   0
  16   0.0001    0 1000   0

Scale:  1 asterisk represents 2 units.


l[6]['Vectors']
               1           2           3           4

  1      0.2576      0.0800     -0.0664      0.4383
  2      0.1816      0.3729      0.2401     -0.1729
  3      0.3717      0.0667     -0.0032     -0.1846
  4      0.0718      0.2321      0.1054      0.4025
  5     -0.0158     -0.2670      0.3756     -0.0695
  6     -0.2316      0.1785     -0.2944      0.3580
  7     -0.2596      0.3093     -0.1880      0.1863
  8     -0.1528      0.4116     -0.1284     -0.4200
  9      0.0317     -0.1108     -0.4861     -0.2513
 10     -0.1832      0.4191      0.2895      0.0453
 11     -0.3304     -0.2500     -0.1372     -0.2305
 12     -0.3153      0.2344      0.1965     -0.1441
 13     -0.1761     -0.2895      0.4655      0.1359
 14     -0.2645     -0.1352     -0.1667      0.2428
 15     -0.3864     -0.1349     -0.0337      0.0868
 16     -0.3454      0.0484      0.1552     -0.1174

 1     allgawp   2       spezp   3      lohnqp
 4      lohnak   5      heizqm   6         eqm
 7      glasqm   8    glasqmak   9         fkp
10     anvermp  11      beinkp  12     beinkak
13    beinkeqm  14     kapkoef  15      rdiffp
16    rentkoef


The first two principal components explain       51.58
percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

     allgawp      0.0728
       spezp      0.1720
      lohnqp      0.1426
      lohnak      0.0590
      heizqm      0.0716
         eqm      0.0855
      glasqm      0.1631
    glasqmak      0.1928
         fkp      0.0133
     anvermp      0.2092
      beinkp      0.1717
     beinkak      0.1544
    beinkeqm      0.1148
     kapkoef      0.0883
      rdiffp      0.1675
    rentkoef      0.1216
```

Übersicht B16b: Vergleich der Hauptkomponentenanalyseergebnisse für Gruppe 8 und Gruppe 13 der 24 Gruppen der Kennzahlenbetriebe

```
*****      Analysis of Subgroup           8       *****

 Total number of units in data =  32

 No outliers were detected
 Number of iterations used =  1

No      Root    %%  Cum    %  Scree Diagram (* represents 1%)
   1   5.4118  338  338  34 **********************************
   2   3.1687  198  536  20 ********************
   3   1.8936  118  655  12 ************
   4   1.3362   84  738   8 ********
   5   1.2970   81  819   8 ********
   6   0.8955   56  875   6 ******
   7   0.6215   39  914   4 ****
   8   0.4767   30  944   3 ***
   9   0.3913   24  968   2 **
  10   0.2780   17  986   2 **
  11   0.0995    6  992   1 *
  12   0.0560    4  995   0
  13   0.0363    2  998   0
  14   0.0319    2 1000   0
  15   0.0052    0 1000   0
  16   0.0008    0 1000   0

Scale:  1 asterisk represents 1 unit.


l[8]['Vectors']
               1           2           3           4

  1      0.1937     -0.3912      0.0844     -0.2782
  2      0.2811      0.1369     -0.2442      0.2024
  3      0.1332      0.3069      0.0253     -0.0672
  4     -0.1084     -0.0563     -0.4429     -0.3841
  5      0.0632      0.4105     -0.1298      0.1149
  6     -0.0935     -0.0992     -0.4549      0.3121
  7     -0.0425     -0.0423     -0.4141      0.4691
  8      0.1491     -0.3536     -0.3344     -0.2418
  9      0.0361      0.2965     -0.0748     -0.4606
 10      0.0305     -0.4469      0.2384      0.2618
 11     -0.3801      0.0272      0.1032      0.0408
 12     -0.3378     -0.2041     -0.2296     -0.2227
 13     -0.3395      0.1329      0.2611      0.0418
 14      0.3272     -0.2462      0.1859      0.0594
 15     -0.4099     -0.0956      0.0188      0.0528
 16     -0.4074     -0.0904      0.0210     -0.0186

 1     allgawp   2       spezp   3     lohnqp
 4      lohnak   5      heizqm   6        eqm
 7      glasqm   8    glasqmak   9        fkp
10     anvermp  11      beinkp  12    beinkak
13    beinkeqm  14     kapkoef  15     rdiffp
16    rentkoef


The first two principal components explain       53.63
percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

     allgawp      0.1906
       spezp      0.0978
      lohnqp      0.1119
      lohnak      0.0149
      heizqm      0.1725
         eqm      0.0186
       glasqm      0.0036
    glasqmak      0.1473
         fkp      0.0892
     anvermp      0.2007
      beinkp      0.1452
     beinkak      0.1557
    beinkeqm      0.1330
     kapkoef      0.1677
      rdiffp      0.1772
    rentkoef      0.1742
```

```
*****      Analysis of Subgroup          13       *****

 Total number of units in data =  24

 No outliers were detected
 Number of iterations used =  1

 No      Root    %%  Cum    %  Scree Diagram (* represents 2%)
  1    5.922  370  370  37 ******************
  2    3.152  197  567  20 **********
  3    2.309  144  711  14 *******
  4    1.299   81  793   8 ****
  5    0.917   57  850   6 ***
  6    0.739   46  896   5 ***
  7    0.499   31  927   3 **
  8    0.355   22  950   2 *
  9    0.320   20  970   2 *
 10    0.229   14  984   1 *
 11    0.163   10  994   1 *
 12    0.066    4  998   0
 13    0.021    1 1000   0
 14    0.005    0 1000   0
 15    0.003    0 1000   0
 16    0.000    0 1000   0

Scale:  1 asterisk represents 2 units.


l[13]['Vectors']
               1           2           3           4

  1      0.2453     -0.3026     -0.0280     -0.0087
  2      0.3471      0.0630     -0.2600      0.0498
  3      0.2495      0.1277      0.4130      0.1579
  4      0.0046     -0.0026     -0.0646      0.7804
  5     -0.0065     -0.3737     -0.1845     -0.1575
  6      0.1020      0.4004      0.2886      0.0925
  7      0.0981      0.4727      0.1624     -0.1240
  8      0.0251      0.2886     -0.4093      0.2884
  9      0.1768      0.1969     -0.4314     -0.0234
 10     -0.0534     -0.3337      0.0621      0.3595
 11     -0.3610      0.2031      0.1523     -0.0011
 12     -0.3664      0.0737     -0.1348      0.254
 13     -0.3630     -0.1436     -0.0046     -0.0337
 14     -0.0636     -0.1867      0.4449      0.1568
 15     -0.3892      0.1006     -0.1007     -0.1035
 16     -0.3877      0.1382     -0.0791     -0.0255

 1     allgawp   2       spezp   3      lohnqp
 4      lohnak   5      heizqm   6         eqm
 7      glasqm   8    glasqmak   9         fkp
10     anvermp  11      beinkp  12     beinkak
13    beinkeqm  14     kapkoef  15      rdiffp
16    rentkoef


The first two principal components explain       56.71
percent of the total variation in the data

The adequacy of fit of the variables in two dimensions is

     allgawp      0.1517
       spezp      0.1244
      lohnqp      0.0786
      lohnak      0.0000
      heizqm      0.1397
         eqm      0.1707
       glasqm      0.2330
    glasqmak      0.0839
         fkp      0.0700
     anvermp      0.1142
      beinkp      0.1716
     beinkak      0.1397
    beinkeqm      0.1524
     kapkoef      0.0389
      rdiffp      0.1616
    rentkoef      0.1694
```

Übersicht B17: Eigenwerte und kanonische Mittelwerte der kanonischen Variablenanalyse der 24 Gruppen der Kennzahlenbetriebe; Grundlage der Analyse ist die Matrix der Summen und Produkte der 24 Gruppen, gewichtet mit den Wichtungsfaktoren nach Ausreißeranalyse (CAMPBELL, 1980)

```
*****  Canonical variate analysis
*****

***  Latent Roots  ***

    lcon['Roots']
              1           2
          4.505       0.829

***  Percentage variation  ***

    lcon['Roots']
              1           2
          77.59       14.28

***  Trace  ***

 lcon['Trace']
       5.806
```

*** Canonical Variate Means ***

	cvacon 1	2
1	-2.331	-1.120
2	-2.420	-0.988
3	-2.406	-0.703
4	-2.110	0.905
5	-2.070	0.942
6	-2.024	1.293
7	-1.003	-0.977
8	-1.043	-1.060
9	-1.030	-0.771
10	-0.956	0.990
11	-0.985	1.139
12	-0.978	1.192
13	0.319	-0.549
14	0.275	-0.710
15	0.208	-0.600
16	0.336	0.781
17	0.340	0.769
18	0.306	0.807
19	3.086	-1.132
20	3.104	-1.138
21	3.225	-0.927
22	3.471	0.304
23	3.443	0.296
24	3.402	0.606

Übersicht B18: Variablen und ihre Skalierung oder Transformation in Gruppierungs- und Segmentierungsanalysen

Kennzahl	Skalierung in modell-begründeter Clustera-naylse	Skalierung für hierar-chische und nicht hierarchische Clu-steranalyse zur Er-rechnung der Proximi-tätsmatrix	Kennzahlen in CART	Skalierung der Kenn-zahlen in CHAID (ordinal skalierte Variablen)
fabswg	-	nominal	nominal	nominal, frei
fregion	-	nominal	nominal	nominal, frei
ak	standardisiert (r)	ordinal	(r)	ordinal, monoton
epertp	standardisiert (r)	ordinal	(r)	ordinal, monoton
eqm	standardisiert (r)	ordinal	(r)	ordinal, monoton
fremdakp	standardisiert (r)	ordinal	(r)	ordinal, monoton
glasqm	standardisiert (r)	ordinal	(r)	ordinal, monoton
glasqmak	standardisiert (r)	ordinal	(r)	ordinal, monoton
anvermp	standardisiert (r)	ordinal	(r)	ordinal, monoton
fkp	standardisiert (r)	ordinal	(r)	ordinal, monoton
netinvp	standardisiert (i)	ordinal	(i)	ordinal, monoton
verm	standardisiert (r)	ordinal	(r)	ordinal, monoton
allgawp	-	-	(r)	-
heizqm	standardisiert (r)	ordinal	(r)	ordinal, monoton
lohnak	standardisiert (r)	ordinal	(r)	ordinal, monoton
lohnqp	-	-	(r)	-
spezp	-	-	(r)	-
beinkak	standardisiert (i)	ordinal	-	-
beinkweqm	standardisiert (i)	ordinal	-	-
beinkp	standardisiert (i)	ordinal	-	-
kapkoef	standardisiert (i)	ordinal	-	-
rentkoef	standardisiert (i)	ordinal	(i)	ordinal, monoton
rdiffp	standardisiert (i)	ordinal	-	-

(i) Intervallskala
(r) Verhältnisskala

Übersicht B19: In CART verwendete, von 1 abweichende Gewichtungen für die Objekte nach multivariater Ausreißerprüfung

Betrieb	Gewicht	Mahalanobis-Distanz	Betrieb	Gewicht	Mahalanobis-Distanz
486	0.000	50.152	728	0.525	6.847
557	0.000	38.299	429	0.528	6.841
501	0.000	27.761	82	0.584	6.724
485	0.000	25.484	656	0.593	6.706
500	0.000	22.651	370	0.599	6.693
386	0.000	17.251	114	0.606	6.679
499	0.000	16.660	786	0.620	6.651
403	0.000	16.221	355	0.623	6.645
387	0.000	15.501	311	0.633	6.624
484	0.000	15.210	770	0.637	6.616
796	0.000	14.368	232	0.663	6.562
417	0.000	13.687	290	0.670	6.548
380	0.000	13.092	258	0.679	6.529
385	0.000	12.834	713	0.686	6.515
882	0.000	12.358	289	0.693	6.500
381	0.000	11.638	759	0.701	6.483
558	0.000	11.623	124	0.707	6.471
401	0.000	11.284	784	0.717	6.450
416	0.000	11.165	819	0.730	6.423
436	0.000	10.750	818	0.730	6.423
743	0.000	10.382	206	0.730	6.422
798	0.001	10.248	163	0.731	6.421
349	0.001	10.236	295	0.742	6.398
2	0.001	10.067	771	0.745	6.390
404	0.002	9.907	283	0.754	6.372
29	0.002	9.869	236	0.762	6.355
379	0.002	9.790	3	0.763	6.351
30	0.003	9.729	147	0.780	6.314
28	0.006	9.456	826	0.799	6.272
371	0.007	9.374	812	0.808	6.252
362	0.009	9.309	1	0.816	6.233
453	0.010	9.227	679	0.818	6.229
64	0.013	9.146	126	0.819	6.226
145	0.015	9.069	194	0.827	6.207
415	0.018	8.982	129	0.831	6.197
634	0.021	8.912	785	0.843	6.169
140	0.025	8.839	117	0.847	6.160
27	0.025	8.835	164	0.853	6.146
881	0.030	8.762	676	0.860	6.128
337	0.037	8.652	678	0.860	6.128
88	0.064	8.370	189	0.877	6.085
146	0.065	8.365	567	0.887	6.057
549	0.072	8.303	712	0.894	6.038
350	0.098	8.132	867	0.896	6.034
207	0.140	7.912	794	0.897	6.031
351	0.142	7.907	571	0.898	6.028
282	0.144	7.894	405	0.902	6.017
452	0.154	7.851	828	0.903	6.014
141	0.160	7.830	697	0.904	6.012
811	0.166	7.804	231	0.905	6.008
222	0.166	7.804	357	0.907	6.002
725	0.171	7.785	112	0.913	5.987
312	0.172	7.781	673	0.915	5.980
451	0.176	7.764	445	0.918	5.972
291	0.196	7.691	150	0.929	5.937
548	0.218	7.613	26	0.939	5.906
547	0.248	7.518	262	0.946	5.882
813	0.250	7.513	276	0.948	5.877
729	0.267	7.461	314	0.956	5.850
261	0.281	7.424	742	0.958	5.840
714	0.317	7.325	274	0.964	5.820
769	0.341	7.264	684	0.966	5.812
674	0.350	7.240	748	0.971	5.791
878	0.373	7.186	762	0.973	5.782
361	0.415	7.086	698	0.976	5.771
460	0.425	7.063	447	0.977	5.766
316	0.453	7.002	595	0.981	5.750
877	0.461	6.983	113	0.985	5.734
705	0.466	6.973	310	0.991	5.705
142	0.471	6.961	869	0.991	5.705
263	0.478	6.946	125	0.992	5.700
727	0.481	6.941	814	0.992	5.700
198	0.483	6.936	827	0.993	5.695
879	0.490	6.922	677	0.996	5.681
726	0.494	6.913	265	0.996	5.680
708	0.507	6.885	139	0.999	5.665

Übersicht B20: Beurteilung der Normalverteilung bei der Kennzahl Rentabilitätskoeffizient im vollen und im eingeschränkten Datensatz inm den Jahren 1992, 1993, 1994

1992

Summary statistics for rentkoef, full sample

Number of values	297
Skewness	0.844
Standard Error of Skewness	0.141
Kurtosis	3.820
Standard Error of Kurtosis	0.282

restrcted cases

jahr	case	rentkoef
92.00	145.0	-0.550
92.00	355.0	2.690
92.00	472.0	1.910
92.00	595.0	2.020
92.00	634.0	2.980
92.00	676.0	2.470
92.00	799.0	1.870

Summary statistics for rentkoef restricted

Number of values	290
Skewness	0.176
Standard Error of Skewness	0.143
Kurtosis	-0.101
Standard Error of Kurtosis	0.285

1993

Summary statistics for rentkoef, full sample

Number of values	297
Skewness	-2.571
Standard Error of Skewness	0.141
Kurtosis	24.485
Standard Error of Kurtosis	0.282

restricted cases

jahr	case	rentkoef
93.00	146.0	0.080
93.00	326.0	1.950
93.00	356.0	2.020
93.00	416.0	0.150
93.00	443.0	0.110
93.00	533.0	1.750
93.00	557.0	-3.120
93.00	611.0	0.150
93.00	623.0	1.970
93.00	656.0	0.100
93.00	677.0	2.010
93.00	788.0	1.770
93.00	800.0	1.910
93.00	881.0	-0.020

Summary statistics for rentkoef restricted

Number of values	283
Skewness	0.142
Standard Error of Skewness	0.145
Kurtosis	-0.264
Standard Error of Kurtosis	0.289

1994

Summary statistics for rentkoef, full sample

Number of values	297
Skewness	0.793
Standard Error of Skewness	0.141
Kurtosis	2.133
Standard Error of Kurtosis	0.282

restricted cases

jahr	case	rentkoef
94.00	222.0	2.650
94.00	297.0	2.040
94.00	357.0	2.600
94.00	393.0	1.940
94.00	417.0	0.090
94.00	486.0	-0.050
94.00	534.0	1.950
94.00	636.0	1.920
94.00	678.0	2.410

Summary statistics for rentkoef restricted

Number of values	288
Skewness	0.229
Standard Error of Skewness	0.144
Kurtosis	-0.096
Standard Error of Kurtosis	0.286

Übersicht B21: Beurteilung der Normalverteilung bei der Kennzahl Rentabilitätskoeffizient im vollen und im eingeschränkten Datendsatz in den Jahren 1992, 1993, 1994

```
Histogram of rentkoef restricted

              - 0.2   2 *
          0.2 - 0.4  10 *****
          0.4 - 0.6  31 ***************
          0.6 - 0.8  42 ********************
          0.8 - 1.0  81 ****************************************
          1.0 - 1.2  59 ******************************
          1.2 - 1.4  38 ******************
          1.4 - 1.6  17 *********
          1.6 - 1.8   9 *****
          1.8 -       1 *

Scale:  1 asterisk represents 2 units.


Stem-and-leaf display for rentkoef restricted
  Number of observations: 290    Minimum: 0.2   Maximum: 1.8
  Stem units: 0.1, leaf digits: 1 (the value 0.1700 is represented
  by  1|7)

   2     1|78
   3     2|668
   5     3|03788
  12     4|002233457889
  18     5|000222367788999999
  17     6|00012244445566679
  27     7|000112234455555666777888999
  39     8|011111222223344445566666777777788889999
  40     9|000012223333344444444555556677777999999999
  39    10|000111223344455556667777778888899999999
  22    11|00112222222223344566899
  24    12|01123445555666777777899
  14    13|13455566678999
   9    14|011234567
   9    15|445667777
   4    16|1378
   4    17|2278
   2    18|01


 ---------------------------------------------------------------------
                                        Test statistic
                                 ---------------------------------
 Type of                         Anderson-      Cramer-
   test        Variate(s)         Darling     von Mises      Watson
 ---------------------------------------------------------------------
 Marginal
              unrestricted       2.368**       0.392**       0.348**
                restricted       0.496         0.080         0.074
 ---------------------------------------------------------------------
```

```
Histogram of rentkoef restricted

              - 0.32   7 ****
          0.32 - 0.48  10 *****
          0.48 - 0.64  30 ***************
          0.64 - 0.80  38 *******************
          0.80 - 0.96  69 **********************************
          0.96 - 1.12  55 ****************************
          1.12 - 1.28  24 ************
          1.28 - 1.44  25 *************
          1.44 - 1.60  17 *********
          1.60 -        8 ****

Scale:  1 asterisk represents 2 units.


  Stem-and-leaf display for rentkoef restricted
  Number of observations: 283    Minimum: 0.2   Maximum: 1.7
  Stem units: 0.1, leaf digits: 1 (the value 0.2200 is represented
  by  2|2)

   6     2|245578
   5     3|23568
  10     4|1457779999
  15     5|122445566666789
  22     6|1122333344455678889999
  21     7|01234566777778889999
  48     8|000000112233333444445555556666667777788888888889
  37     9|000122223333333444456666666777788888999
  34    10|00111222233444556667777788888999999
  23    11|0011111222244455568899
  17    12|01233333456799999
  10    13|1233467799
  21    14|00011223345556788899 9
   6    15|235599
   5    16|14477
   3    17|023


 ---------------------------------------------------------------------
                                        Test statistic
                                 ---------------------------------
 Type of                         Anderson-      Cramer-
   test        Variate(s)         Darling     von Mises      Watson
 ---------------------------------------------------------------------
 Marginal
              unrestricted       3.616**       0.628**       0.623**
                restricted       0.865*        0.146*        0.138*
 ---------------------------------------------------------------------
```

```
Histogram of rentkoef restricted

              - 0.2   0
          0.2 - 0.4   8 ****
          0.4 - 0.6  28 **************
          0.6 - 0.8  46 ************************
          0.8 - 1.0  78 ****************************************
          1.0 - 1.2  60 ******************************
          1.2 - 1.4  35 ******************
          1.4 - 1.6  23 ************
          1.6 - 1.8   8 ****
          1.8 -       2 *

Scale:  1 asterisk represents 2 units.


Stem-and-leaf display for rentkoef restricted
  Number of observations: 288    Minimum: 0.2   Maximum: 1.8
  Stem units: 0.1, leaf digits: 1 (the value 0.2100 is represented
  by  2|1)

   4     2|1239
   4     3|0568
   8     4|23445799
  19     5|1222244455555778899
  13     6|0234455567799
  31     7|00111123445666666667778899999999
  37     8|000111122223333444455555666666777888999
  40     9|0000111112223344445555556666667777778899999
  37    10|000011222222334444555556666677777888899
  24    11|012222333333344555778899
  18    12|000111222334556677
  20    13|01112233333445557899
  14    14|12355667889999
   8    15|23334568
   4    16|0257
   5    17|34589
   2    18|34


 ---------------------------------------------------------------------
                                        Test statistic
                                 ---------------------------------
 Type of                         Anderson-      Cramer-
   test        Variate(s)         Darling     von Mises      Watson
 ---------------------------------------------------------------------
 Marginal
              unrestricted       2.182**       0.382**       0.323**
                restricted       0.699?        0.121?        0.107?
 ---------------------------------------------------------------------
```

Übersicht B22: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1992, Verwendung der Gewichtung nach Ausreißertests

```
Regression tree:
tree(formula = rentkoef.full ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp + fabswg +
      fregion, weights = weights, subset = fjahr ==
      "92")
Variables actually used in tree construction:
[1] "lohnqp"   "spezp"    "allgawp"  "epertp"
[5] "anvermp"  "fremdakp" "fabswg"   "glasqmak"
Number of terminal nodes:  29
Residual mean deviance:  0.01609 = 4.313 / 268
Distribution of residuals:
    Min. 1st Qu.    Median      Mean 3rd Qu.
 -0.8806 -0.0911 0.0004352 -0.004082 0.09595
   Max.
 0.6676

summary(ftree.92.pruned7)

Regression tree:
snip.tree(tree = ftree.92, nodes = c(14, 17, 15, 6,
      16, 9, 5))
Variables actually used in tree construction:
[1] "lohnqp"  "spezp"   "allgawp"
Number of terminal nodes:  7
Residual mean deviance:  0.047 = 13.63 / 290
Distribution of residuals:
   Min. 1st Qu.   Median      Mean 3rd Qu.  Max.
 -1.143 -0.1482 -0.00582 -0.005566  0.1318 1.288


lohnqp[1:297]
   Min. 1st Qu. Median Mean 3rd Qu. Max.
 13.39   26.37  32.48 34.7   40.24  117

spezp[1:297]
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 7.68   25.18  33.49 33.19   39.76 84.2

allgawp[1:297]
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 7.13   19.82  25.33 26.23   32.17 49.31
```

```
      var  n       dev      yval
 5 <leaf> 55 3.2034190 0.9482249
 6 <leaf> 68 2.3018574 0.9758199
 9 <leaf> 34 2.1237838 1.1364736
14 <leaf> 47 1.4713238 0.8507095
15 <leaf> 46 1.4394617 0.5931755
16 <leaf> 21 2.2150736 1.6917334
17 <leaf> 26 0.8744454 1.3349647

node number: 5
   root
   lohnqp<31.63
   spezp>37.175

 node number: 6
   root
   lohnqp>31.63
   spezp<28.285

 node number: 9
   root
   lohnqp<31.63
   spezp<37.175
   allgawp>29.735

 node number: 14
   root
   lohnqp>31.63
   spezp>28.285
   allgawp<23.995

 node number: 15
   root
   lohnqp>31.63
   spezp>28.285
   allgawp>23.995

 node number: 16
   root
   lohnqp<31.63
   spezp<37.175
   allgawp<29.735
   spezp<27.82

 node number: 17
   root
   lohnqp<31.63
   spezp<37.175
   allgawp<29.735
   spezp>27.82
```

Übersicht B23: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1993, Verwendung der Gewichtung nach Ausreißertests

```
Regression tree:
tree(formula = rentkoef.full ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp + fabswg +
      fregion, weights = weights, subset = fjahr ==
      "93")
Variables actually used in tree construction:
 [1] "lohnqp"   "spezp"    "allgawp"  "lohnak"
 [5] "glasqmak" "epertp"   "netinvp"  "glasqm"
 [9] "anvermp"  "verm"     "ak"
Number of terminal nodes:  32
Residual mean deviance:  0.02171 = 5.752 / 265
Distribution of residuals:
   Min. 1st Qu.   Median     Mean 3rd Qu.   Max.
 -3.361 -0.1011 -0.01208 -0.02439 0.08823 0.4382

summary(ftree.93.pruned7)

Regression tree:
snip.tree(tree = ftree.93, nodes = c(13, 15, 12, 14,
      10, 4))
Variables actually used in tree construction:
[1] "lohnqp"  "spezp"   "netinvp" "allgawp"
Number of terminal nodes:  7
Residual mean deviance:  0.05104 = 14.8 / 290
Distribution of residuals:
   Min. 1st Qu.   Median     Mean 3rd Qu.   Max.
 -3.739 -0.1693 -0.01974 -0.03525  0.1381 0.5881


lohnqp[298:594]
  Min. 1st Qu. Median  Mean 3rd Qu. Max.
 11.91    26.9  32.42 35.09   40.39  166

spezp[298:594]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 10.81   24.87  32.85 34.03   39.81 248.2

allgawp[298:594]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 10.31   20.97  25.92 26.72   31.89 64.48

netinvp[298:594]
   Min. 1st Qu. Median   Mean 3rd Qu.  Max.
 -151.6  -14.97  -6.19 -2.454    8.23 88.62
```

```
      var  n       dev      yval
 4 <leaf> 39 3.2040579 1.4621080
10 <leaf> 39 2.8341996 1.1365090
11 <leaf>  8 0.7980908 0.5902510
12 <leaf> 21 1.4660561 1.2420409
13 <leaf> 57 1.3328610 0.9497415
14 <leaf> 87 3.8170427 0.9018784
15 <leaf> 46 1.3500855 0.6193360

node number: 4
   root
   lohnqp<28.07
   spezp<33.58

 node number: 10
   root
   lohnqp<28.07
   spezp>33.58
   netinvp<23.66

 node number: 11
   root
   lohnqp<28.07
   spezp>33.58
   netinvp>23.66

 node number: 12
   root
   lohnqp>28.07
   spezp<27.265
   lohnqp<33.46

 node number: 13
   root
   lohnqp>28.07
   spezp<27.265
   lohnqp>33.46

 node number: 14
   root
   lohnqp>28.07
   spezp>27.265
   allgawp<27.995

 node number: 15
   root
   lohnqp>28.07
   spezp>27.265
   allgawp>27.995
```

Übersicht B24: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1994, Verwendung der Gewichtung nach Ausreißertests

```
Regression tree:
tree(formula = rentkoef.full ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp + fabswg +
      fregion, weights = weights, subset = fjahr ==
      "94")
Variables actually used in tree construction:
[1] "lohnqp"   "spezp"    "allgawp"  "glasqm"
[5] "fremdakp" "netinvp"  "epertp"   "ak"
[9] "eqm"
Number of terminal nodes:  31
Residual mean deviance:  0.01932 = 5.14 / 266
Distribution of residuals:
    Min.  1st Qu.  Median      Mean 3rd Qu.  Max.
 -0.7008 -0.09143 -0.0055 -0.005761 0.09091 0.574

summary(ftree.94.pruned7)

Regression tree:
snip.tree(tree = ftree.94, nodes = c(8, 13, 7, 11,
      12))
Variables actually used in tree construction:
[1] "lohnqp"   "spezp"    "allgawp"  "fremdakp"
Number of terminal nodes:  7
Residual mean deviance:  0.0474 = 13.74 / 290
Distribution of residuals:
    Min. 1st Qu.  Median     Mean 3rd Qu.   Max.
 -0.7332 -0.1632 -0.0256 -0.01377  0.1394 0.7205


lohnqp[595:891]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 11.25   26.63  32.26 35.44   40.09 242.7

spezp[595:891]
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 6.08   24.34  31.31 31.89   38.45 67.12

allgawp[595:891]
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 7.95   21.29  25.92 26.77   31.81 56.27

fremdakp[595:891]
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
    0   49.32  68.05 61.89    77.9 94.51
```

```
      var  n       dev      yval
 7 <leaf> 58 1.8191220 0.6831542
 8 <leaf> 11 1.0165926 1.9295484
 9 <leaf>  5 0.5973200 1.3460000
10 <leaf>  9 0.3686927 1.6208248
11 <leaf> 72 4.0385654 1.1706246
12 <leaf> 73 3.7463363 1.0855977
13 <leaf> 69 2.1582423 0.8408787

 node number: 7
   root
   lohnqp>28.29
   lohnqp>43.46

 node number: 8
   root
   lohnqp<28.29
   spezp<22.88
   allgawp<34.34

 node number: 9
   root
   lohnqp<28.29
   spezp<22.88
   allgawp>34.34

 node number: 10
   root
   lohnqp<28.29
   spezp>22.88
   fremdakp<32.415

 node number: 11
   root
   lohnqp<28.29
   spezp>22.88
   fremdakp>32.415

 node number: 12
   root
   lohnqp>28.29
   lohnqp<43.46
   spezp<31.235

 node number: 13
   root
   lohnqp>28.29
   lohnqp<43.46
   spezp>31.235
```

Übersicht B25: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1992, um Extremwerte verkleinerter Datensatz

```
Regression tree:
tree(formula = rentkoef.rest ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp +
      f.abswg.rest + f.region.rest, subset =
      f.jahr.rest == "92")
Variables actually used in tree construction:
[1] "lohnqp"   "spezp"    "allgawp"  "verm"
[5] "glasqmak" "epertp"   "anvermp"  "fremdakp"
Number of terminal nodes:  29
Residual mean deviance:  0.01742 = 4.548 / 261
Distribution of residuals:
   Min.  1st Qu.   Median       Mean 3rd Qu.
 -0.404 -0.07656 0.008063 2.335e-017   0.091
   Max.
 0.2817

summary(ftree.92.pruned7)

Regression tree:
snip.tree(tree = ftree.92.rest, nodes = c(8, 14, 15,
      6, 19, 5))
Variables actually used in tree construction:
[1] "lohnqp"  "spezp"   "allgawp"
Number of terminal nodes:  7
Residual mean deviance:  0.04365 = 12.35 / 283
Distribution of residuals:
    Min. 1st Qu.   Median        Mean 3rd Qu.
 -0.5697 -0.1276 0.001017 -3.331e-017  0.1346
   Max.
 0.5401


lohnqp[1:290]
  Min. 1st Qu. Median  Mean 3rd Qu. Max.
 13.39    26.5  32.66 34.92   40.26  117

spezp[1:290]
 Min. 1st Qu. Median Mean 3rd Qu.  Max.
 7.68   25.38  33.72 33.3    39.9 65.93

allgawp[1:290]
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 7.13   20.02  25.47 26.32   32.18 49.31
```

```
      var  n       dev      yval
 5 <leaf> 59 3.3245932 0.9496610
 6 <leaf> 74 2.9268986 0.9298649
 8 <leaf> 48 2.2121000 1.3825000
14 <leaf> 41 1.2768390 0.7712195
15 <leaf> 28 0.8765429 0.4885714
18 <leaf>  7 0.3017714 1.4942857
19 <leaf> 33 1.4343879 0.9893939


 node number: 5
   root
   lohnqp<32.87
   spezp>37.665

 node number: 6
   root
   lohnqp>32.87
   spezp<32.365

 node number: 8
   root
   lohnqp<32.87
   spezp<37.665
   allgawp<30.23

node number: 14
   root
   lohnqp>32.87
   spezp>32.365
   lohnqp<44.32

 node number: 15
   root
   lohnqp>32.87
   spezp>32.365
   lohnqp>44.32

 node number: 18
   root
   lohnqp<32.87
   spezp<37.665
   allgawp>30.23
   lohnqp<21.99

 node number: 19
   root
   lohnqp<32.87
   spezp<37.665
   allgawp>30.23
   lohnqp>21.99
```

Übersicht B26: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1993, um Extremwerte verkleinerter Datensatz

```
Regression tree:
tree(formula = rentkoef.rest ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp +
      f.abswg.rest + f.region.rest, subset =
      f.jahr.rest == "93")
Variables actually used in tree construction:
 [1] "lohnqp"   "spezp"    "allgawp"  "heizqm"
 [5] "netinvp"  "epertp"   "verm"     "ak"
 [9] "fremdakp" "anvermp"
Number of terminal nodes:  30
Residual mean deviance:  0.01828 = 4.625 / 253
Distribution of residuals:
    Min.  1st Qu. Median        Mean 3rd Qu.
 -0.6033 -0.07806   0.01 -2.079e-017  0.0819
   Max.
 0.3967

summary(ftree.93.pruned7)

Regression tree:
snip.tree(tree = ftree.93.rest, nodes = c(24, 25,
      13, 9, 5))
Variables actually used in tree construction:
[1] "lohnqp"  "spezp"   "allgawp"
Number of terminal nodes:  7
Residual mean deviance:  0.04621 = 12.75 / 276
Distribution of residuals:
   Min. 1st Qu.   Median        Mean 3rd Qu.
 -0.833 -0.1263 -0.00451 -6.748e-017  0.1287
   Max.
 0.7803


lohnqp[291:578]
  Min. 1st Qu. Median Mean 3rd Qu. Max.
 11.91   27.33  32.69 35.1   40.29  166

spezp[291:578]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 10.81   25.44  32.85 33.07   39.76 63.51

allgawp[291:578]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 10.31   21.04  25.96 26.54   31.92 60.88
```

```
      var  n       dev      yval
 5 <leaf> 59 4.5655932 0.9196610
 7 <leaf> 12 0.2311667 0.4216667
 8 <leaf> 14 0.1419500 1.5250000
 9 <leaf> 71 4.3826789 1.1929577
13 <leaf> 24 0.8040500 0.6375000
24 <leaf> 51 1.0918627 0.9845098
25 <leaf> 52 1.5355923 0.7896154


 node number: 5
   root
   lohnqp<32.92
   spezp>37.76

 node number: 7
   root
   lohnqp>32.92
   lohnqp>57.485

 node number: 8
   root
   lohnqp<32.92
   spezp<37.76
   allgawp<20.595

 node number: 9
   root
   lohnqp<32.92
   spezp<37.76
   allgawp>20.595

node number: 13
   root
   lohnqp>32.92
   lohnqp<57.485
   spezp>39.95

node number: 24
   root
   lohnqp>32.92
   lohnqp<57.485
   spezp<39.95
   allgawp<26.2

 node number: 25
   root
   lohnqp>32.92
   lohnqp<57.485
   spezp<39.95
   allgawp>26.2
```

Übersicht B27: Beschreibung des vollen und des auf sieben Terminalknoten gestutzten Regressionsbaums 1994, um Extremwerte verkleinerter Datensatz

```
Regression tree:
tree(formula = rentkoef.rest ~ ak + epertp + eqm +
      fremdakp + glasqm + glasqmak + anvermp +
      fkp + fkp + netinvp + verm + allgawp +
      heizqm + lohnak + lohnqp + spezp +
      f.abswg.rest + f.region.rest, subset =
      f.jahr.rest == "94")
Variables actually used in tree construction:
[1] "lohnqp"   "spezp"    "allgawp"  "heizqm"
[5] "netinvp"  "anvermp"  "ak"       "fremdakp"
[9] "epertp"
Number of terminal nodes:  32
Residual mean deviance:  0.01665 = 4.262 / 256
Distribution of residuals:
    Min.  1st Qu.   Median       Mean 3rd Qu.
 -0.3567 -0.07737 0.005934 1.793e-017 0.07649
   Max.
 0.3075

summary(ftree.94.pruned7)

Regression tree:
snip.tree(tree = ftree.94.rest, nodes = c(24, 25,
      13, 8, 7, 5))
Variables actually used in tree construction:
[1] "lohnqp"  "spezp"   "allgawp"
Number of terminal nodes:  7
Residual mean deviance:  0.04626 = 13 / 281
Distribution of residuals:
    Min. 1st Qu.    Median        Mean 3rd Qu.
 -0.7106 -0.1331 -0.005625 -1.442e-016  0.1296
   Max.
 0.6094


lohnqp[579:861]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 11.25      27  32.36 34.86   40.16 102.6

spezp[579:861]
  Min. 1st Qu. Median Mean 3rd Qu.  Max.
 6.08   24.42  31.45   32   38.46 67.12

allgawp[579:861]
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 7.95   21.35  25.88 26.76   31.68 56.27
```

```
var  n       dev      yval
 5 <leaf> 80 4.7194688 1.0606250
 7 <leaf> 52 2.3973231 0.6523077
 8 <leaf> 39 2.4580359 1.4312821
 9 <leaf>  9 0.1170222 1.0644444
13 <leaf> 30 1.2721367 0.7376667
24 <leaf> 30 0.9526967 1.1436667
25 <leaf> 48 1.0815812 0.9056250


 node number: 5
   root
   lohnqp<31.085
   spezp>30.175

 node number: 7
   root
   lohnqp>31.085
   lohnqp>44.105

 node number: 8
   root
   lohnqp<31.085
   spezp<30.175
   allgawp<36.435

 node number: 9
   root
   lohnqp<31.085
   spezp<30.175
   allgawp>36.435

 node number: 13
   root
   lohnqp>31.085
   lohnqp<44.105
   allgawp>29.775

node number: 24
   root
   lohnqp>31.085
   lohnqp<44.105
   allgawp<29.775
   spezp<27.455

 node number: 25
   root
   lohnqp>31.085
   lohnqp<44.105
   allgawp<29.775
   spezp>27.455
```

Übersicht B28: Minima, Maxima und Quartile für die Prediktorvariablen, pro Jahr 297 Werte

	Minimum	Perzen-tile 25	Median	Perzen-tile 75	Maximum
jahr					
1992					
anzahl arbeitskräfte	1,00	3,03	4,61	6,86	17,50
anlagevermögen	2,16	37,73	46,86	55,67	93,61
anteil eigenproduktion	61,76	90,97	96,83	98,66	100,00
eqm	11000,00	71424,00	116000,0	179388,0	485000,0
anteil fremdkapital	,00	54,92	110,40	180,40	1801,13
anteil fremd-ak	,00	50,00	66,67	78,57	100,00
glasfläche	450,00	3260,00	4665,00	7462,00	18817,00
glasfläche je ak	68,10	734,51	1063,79	1493,51	4572,82
heizung je qm	1,24	5,55	8,47	13,25	34,27
lohn je ak	,00	24146,67	31808,61	38911,00	104592,9
nettoinvestitionen	-318,01	-13,21	-3,76	12,39	205,85
vermögen	67,00	374,00	573,00	903,00	3101,00
1993					
anzahl arbeitskräfte	1,00	2,98	4,69	6,72	18,00
anlagevermögen	1,21	37,73	45,14	53,72	81,56
anteil eigenproduktion	61,54	90,33	96,81	98,61	100,00
eqm	12000,00	75880,00	120000,0	186284,0	528200,0
anteil fremdkapital	,00	59,17	108,24	186,89	2417,20
anteil fremd-ak	,00	49,81	66,67	78,66	100,00
glasfläche	450,00	3350,00	4900,00	7500,00	20000,00
glasfläche je ak	61,78	733,33	1140,07	1542,06	6655,56
heizung je qm	,00	6,11	8,94	12,60	37,35
lohn je ak	,00	26140,69	33303,26	40299,94	114511,0
nettoinvestitionen	-151,56	-14,97	-6,19	8,23	88,62
vermögen	83,00	389,00	606,00	936,00	3811,00
1994					
anzahl arbeitskräfte	1,02	3,00	4,60	6,91	18,22
anlagevermögen	,00	36,12	44,00	53,59	78,16
anteil eigenproduktion	62,69	89,64	96,84	98,82	100,00
eqm	12000,00	80000,00	120000,0	187200,0	469886,0
anteil fremdkapital	,00	61,52	109,13	193,25	3019,67
anteil fremd-ak	,00	49,32	68,05	77,90	94,51
glasfläche	450,00	3500,00	5000,00	7767,00	20000,00
glasfläche je ak	56,73	755,56	1162,55	1616,67	4654,97
heizung je qm	,42	4,52	7,18	10,84	33,19
lohn je ak	,00	25806,52	33660,74	41264,86	78459,00
nettoinvestitionen	-86,37	-15,51	-6,98	6,10	241,64
vermögen	80,00	386,00	611,00	933,00	3526,00
1992-1994					
anzahl arbeitskräfte	1,00	3,00	4,63	6,84	18,22
anlagevermögen	,00	36,80	45,46	54,43	93,61
anteil eigenproduktion	61,54	90,24	96,84	98,66	100,00
eqm	11000,00	75880,00	118000,0	184728,0	528200,0
anteil fremdkapital	,00	58,46	109,63	186,58	3019,67
anteil fremd-ak	,00	50,00	66,67	78,57	100,00
glasfläche	450,00	3350,00	4900,00	7500,00	20000,00
glasfläche je ak	56,73	745,45	1128,43	1538,46	6655,56
heizung je qm	,00	5,39	8,27	12,22	37,35
lohn je ak	,00	25192,51	32908,61	40267,63	114511,0
nettoinvestitionen	-318,01	-14,80	-5,81	9,53	241,64
vermögen	67,00	383,00	596,00	922,00	3811,00

rentabilitätskoeffizient 1992

Percentile	Value	Percentile	Value	Percentile	Value	Percentile	Value
12,50	,590	37,50	,870	62,50	1,070	87,50	1,388
Valid cases	297	Missing cases	0				

rentabilitätskoeffizient 1993

Percentile	Value	Percentile	Value	Percentile	Value	Percentile	Value
12,50	,560	37,50	,850	62,50	1,050	87,50	1,418
Valid cases	297	Missing cases	0				

rentabilitätskoeffizient 1994

Percentile	Value	Percentile	Value	Percentile	Value	Percentile	Value
12,50	,593	37,50	,870	62,50	1,070	87,50	1,445
Valid cases	297	Missing cases	0				

Übersicht B29: Ergebnisse der Segmentierung durch CHAID für die Kennzahlen der Jahre 1992, 1993 und 1994

Segmentierung 1992

f_rentko levels=4 (Dependent) rentabilitätskoeffizient

level	value	symbol: label	frequency	scores
1)	1	1: sehr gering	76	0.59
2)	2	2: gering	74	0.87
3)	3	3: hoch	73	1.07
4)	4	4: sehr hoch	74	1.39

id	count	score	vars...			
-1-	28	0.70	f_eqm=12	f_frmdak=1	f_glqmak=12	
-2-	8	0.72	f_eqm=12	f_frmdak=1	f_glqmak=34	f_anverm=1
-3-	18	1.00	f_eqm=12	f_frmdak=1	f_glqmak=34	f_anverm=2-4
-4-	99	0.94	f_eqm=12	f_frmdak=2-4		
-5-	51	0.96	f_eqm=34	region=ü		
-6-	59	1.07	f_eqm=34	region=r	f_netinv=1-3	
-7-	34	1.22	f_eqm=34	region=r	f_netinv=4	

Id	size	% of all	score index		Cum: size	% of all	score index
7	34	11.4	1.22	125	34	11.4	1.22 125
6	59	19.9	1.07	110	93	31.3	1.13 115
3	18	6.1	1.00	102	111	37.4	1.10 113
5	51	17.2	0.96	98	162	54.5	1.06 108
4	99	33.3	0.94	96	261	87.9	1.01 104
2	8	2.7	0.72	74	269	90.6	1.01 103
1	28	9.4	0.70	72	297	100.0	0.98 100

Segmenmtierung 1993

f_rentko levels=4 (Dependent) rentabilitätskoeffizient

level	value	symbol: label	frequency	scores
1)	1	1: sehr gering	74	0.56
2)	2	2: gering	80	0.85
3)	3	3: hoch	72	1.05
4)	4	4: sehr hoch	71	1.42

id	count	score	vars...				
-1-	12	0.63	f_fkp=12	region=ü	f_ak=1		
-2-	11	0.81	f_fkp=12	region=ü	f_ak=2		
-3-	29	1.10	f_fkp=12	region=ü	f_ak=34	f_lohnak=1-3	
-4-	16	0.90	f_fkp=12	region=ü	f_ak=34	f_lohnak=4	
-5-	15	1.16	f_fkp=12	region=r	f_frmdak=1	f_verm=12	
-6-	14	0.81	f_fkp=12	region=r	f_frmdak=1	f_verm=34	
-7-	19	1.31	f_fkp=12	region=r	f_frmdak=2-4	f_ak=12	
-8-	23	1.11	f_fkp=12	region=r	f_frmdak=2-4	f_ak=3	
-9-	11	1.30	f_fkp=12	region=r	f_frmdak=2-4	f_ak=4	
-10-	147	0.88	f_fkp=34				

Id	size	% of all	score index		Cum: size	% of all	score index
7	19	6.4	1.31	136	19	6.4	1.31 136
9	11	3.7	1.30	135	30	10.1	1.31 136
5	15	5.1	1.16	120	45	15.2	1.26 131
8	23	7.7	1.11	115	68	22.9	1.21 125
3	29	9.8	1.10	115	97	32.7	1.18 122
4	16	5.4	0.90	93	113	38.0	1.14 118
10	147	49.5	0.88	92	260	87.5	0.99 103
2	11	3.7	0.81	85	271	91.2	0.98 102
6	14	4.7	0.81	84	285	96.0	0.98 101
1	12	4.0	0.63	66	297	100.0	0.96 100

Segmentierung 1994

f_rentko levels=4 (Dependent) rentabilitätskoeffizient

level	value	symbol: label	frequency	scores
1)	1	1: sehr gering	69	0.59
2)	2	2: gering	72	0.87
3)	3	3: hoch	76	1.07
4)	4	4: sehr hoch	80	1.45

id	count	score	vars...				
-1-	68	0.87	f_eqm=1				
-2-	24	1.26	f_eqm=23	f_fkp=12	f_heizqm=1		
-3-	44	1.06	f_eqm=23	f_fkp=12	f_heizqm=2-4		
-4-	27	0.97	f_eqm=23	f_fkp=34	f_netinv=1		
-5-	21	0.78	f_eqm=23	f_fkp=34	f_netinv=2		
-6-	25	1.06	f_eqm=23	f_fkp=34	f_netinv=34	f_heizqm=1-3	
-7-	10	0.80	f_eqm=23	f_fkp=34	f_netinv=34	f_heizqm=4	
-8-	13	1.12	f_eqm=4 region=ü	f_lohnak=1-3			
-9-	13	0.85	f_eqm=4 region=ü	f_lohnak=4			
-10-	52	1.19	f_eqm=4 region=r				

Id	size	% of all	score index		Cum: size	% of all	score index
2	24	8.1	1.26	125	24	8.1	1.26 125
10	52	17.5	1.19	118	76	25.6	1.21 120
8	13	4.4	1.12	111	89	30.0	1.20 119
6	25	8.4	1.06	105	114	38.4	1.17 116
3	44	14.8	1.06	105	158	53.2	1.14 112
4	27	9.1	0.97	96	185	62.3	1.11 110
1	68	22.9	0.87	86	253	85.2	1.05 104
9	13	4.4	0.85	84	266	89.6	1.04 103
7	10	3.4	0.80	79	276	92.9	1.03 102
5	21	7.1	0.78	77	297	100.0	1.01 100

Übersicht B30: Direkte Beziehungen nach Screening in den Jahren 1992, 1993 und 1994 zwischen den sechs ausgewählten Erfolgskennzahlen und den 14 übrigen ausgewählten Kennzahlen

direkte Beziehung zwischen...	Jahr 1992 und	1993 und	1994 und
beinkak	abswg - eqm fkp glasqmak lohnak netinvp	- - eqm fkp glasqmak lohnak -	- epertp eqm fkp glasqmak lohnak -
beinkeqm	region ak eqm fkp fremdakp glasqmak heizqm lohnak	region ak eqm fkp - glasqmak heizqm lohnak	region ak - fkp - glasqmak heizqm lohnak
beinkp	eqm fkp heizqm lohnak	- fkp heizqm lohnak	- fkp heizqm lohnak
kapkoef	- ak anvermp - - lohnak	- - anvermp fkp fremdakp lohnak	region - anvermp fkp - -
rdiffp	region epertp eqm fkp fremdakp - - -	region epertp eqm fkp fremdakp - - -	region epertp eqm fkp - glasqmak heizqm netinvp
rentkoef	- - eqm fkp - glasqmak heizqm netinvp	region - eqm fkp fremdakp - heizqm -	region epertp eqm fkp - glasqmak heizqm netinvp

Übersicht B31: Eliminierte Verbindungen nach Rückwärts-Elimination oder EH-Algorithmus für die Analyse von sechs graphischen Modellen 1993

Kennzahlen

```
The main graph.

Variables   abcde
 beinkak: a *++++
     eqm: b +*+++
     fkp: c ++*++
glasqmak: d +++*+
  lohnak: e ++++*
```

Rückwärts-Elimination

```
Model search finished.

bc:      eqm &      fkp   eliminated
cd:      fkp & glasqmak   eliminated
ce:      fkp &   lohnak   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may be removed from all
identified models
  Evidence
 Excl. Incl.       edge
 0.325 0.142  bc -      eqm &      fkp

The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2
 0.142 0.000 * de - glasqmak &   lohnak   +   -
 0.090 0.336   cd -      fkp & glasqmak   -   +
 0.191 0.083   ce -      fkp &   lohnak   -   +
```

Kennzahlen

```
The main graph.

Variables   abcdef
beinkeqm: a *+++++
      ak: b +*++++
     fkp: c ++*+++
glasqmak: d +++*++
  heizqm: e ++++*+
  lohnak: f +++++*
```

Rückwärts-Elimination

```
Model search finished.

bc:       ak &      fkp   eliminated
be:       ak &   heizqm   eliminated
ce:      fkp &   heizqm   eliminated
cf:      fkp &   lohnak   eliminated
ef:   heizqm &   lohnak   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2   3
 0.168 0.008 * bf -       ak &   lohnak   +   -   +
 0.078 0.000 * de - glasqmak &   heizqm   +   -   -
 0.066 0.328   bc -       ak &      fkp   -   -   +
 0.112 0.110   be -       ak &   heizqm   -   +   +
 0.106 0.104   cf -      fkp &   lohnak   -   +   -
 0.462 0.217   ef -   heizqm &   lohnak   -   +   +
```

Kennzahlen

```
The main graph.

Variables   abcd
  beinkp: a *+++
     fkp: b +*++
  heizqm: c ++*+
  lohnak: d +++*
```

Rückwärts-Elimination

```
Model search finished.

bc:      fkp &   heizqm   eliminated
cd:   heizqm &   lohnak   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2
 0.062 0.019 * bd -      fkp &   lohnak   +   -
 0.285 0.325   bc -      fkp &   heizqm   -   +
 0.060 0.114   cd -   heizqm &   lohnak   -   +
```

Kennzahlen

```
The main graph.

Variables   abcde
 kapkoef: a *++++
 anvermp: b +*+++
     fkp: c ++*++
fremdakp: d +++*+
  lohnak: e ++++*
```

Rückwärts-Elimination

```
Model search finished.

be:  anvermp &   lohnak   eliminated
cd:      fkp & fremdakp   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2   3
 0.068 0.004 * be -  anvermp &   lohnak   -   +   +
 0.060 0.004 * ce -      fkp &   lohnak   +   +   -
 0.060 0.017 * de - fremdakp &   lohnak   +   -   +
 0.250 0.384   cd -      fkp & fremdakp   -   +   +
```

Kennzahlen

```
The main graph.

Variables   abcdef
  rdiffp: a *+++++
  region: b +*++++
  epertp: c ++*+++
     eqm: d +++*++
     fkp: e ++++*+
fremdakp: f +++++*
```

Rückwärts-Elimination

```
Model search finished.

ac:   rdiffp &   epertp   eliminated
ad:   rdiffp &      eqm   eliminated
bc:   region &   epertp   eliminated
be:   region &      fkp   eliminated
ce:   epertp &      fkp   eliminated
ef:      fkp & fremdakp   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2   3   4   5   6   7
 0.092 0.000 * ab -   rdiffp &   region   +   +   +   +   -   -   +
 0.268 0.002 * ac -   rdiffp &   epertp   +   +   -   -   +   +   +
 0.242 0.001 * ad -   rdiffp &      eqm   -   -   +   +   +   +   +
 0.068 0.000 * ae -   rdiffp &      fkp   +   +   +   +   +   +   -
 0.396 0.001 * ce -   epertp &      fkp   +   -   -   +   +   -   +
 0.222 0.006 * cf -   epertp & fremdakp   -   +   +   +   -   +   -
 0.078 0.036 * de -      eqm &      fkp   -   +   +   -   -   +   +
 0.258 0.311   be -   region &      fkp   -   -   -   -   +   +   +
 0.360 0.168   ef -      fkp & fremdakp   +   -   -   +   +   -   +
```

Kennzahlen

```
The main graph.

Variables   abcdef
rentkoef: a *+++++
  region: b +*++++
     eqm: c ++*+++
     fkp: d +++*++
fremdakp: e ++++*+
  heizqm: f +++++*
```

Rückwärts-Elimination

```
Model search finished.

ac: rentkoef &      eqm   eliminated
bd:   region &      fkp   eliminated
cd:      eqm &      fkp   eliminated
de:      fkp & fremdakp   eliminated
df:      fkp &   heizqm   eliminated
```

Akzeptierte Modelle im EH-Algorithmus

```
The following edges may be removed from all
identified models
  Evidence
 Excl. Incl.       edge
 0.194 0.314  df -      fkp &   heizqm

The following edges may only be removed (-) for a
subset of models
  Evidence                                Model no.
 Excl. Incl.          edge                1   2   3
 0.136 0.000 * ab - rentkoef &   region   +   -   -
 0.074 0.108   ac - rentkoef &      eqm   -   +   +
 0.152 0.452   bd -   region &      fkp   -   +   +
 0.074 0.066   cd -      eqm &      fkp   -   +   -
 0.300 0.437   de -      fkp & fremdakp   -   -   +
```

Übersicht B32: Modellsuche graphischer Modelle bei der Analyse von sechs Erfolgskennzahlen, 1992 bis 1994

a) EH-Algorithmus 1992

Step 1: Tests for conditional independence
The minimal model will be defined by removal of the following edges:

ab - beinkak & beinkeqm p = 0.0660
ae - beinkak & rdiffp p = 0.1279
af - beinkak & rentkoef p = 0.2860
bc - beinkeqm & beinkp p = 0.0580
be - beinkeqm & rdiffp p = 0.0938
bf - beinkeqm & rentkoef p = 0.1940
ce - beinkp & rdiffp p = 0.0800
de - kapkoef & rdiffp p = 0.3514
df - kapkoef & rentkoef p = 0.6820

Step 2: Examination of the adequacy of the minimal model
Evidence found in the following cases:

ab - beinkak & beinkeqm p = 0.0019 (p = 0.0660)
ae - beinkak & rdiffp p = 0.0000 (p = 0.1279)
af - beinkak & rentkoef p = 0.0000 (p = 0.2860)
be - beinkeqm & rdiffp p = 0.0003 (p = 0.0938)
bf - beinkeqm & rentkoef p = 0.0020 (p = 0.1940)
ce - beinkp & rdiffp p = 0.0040 (p = 0.0800)
de - kapkoef & rdiffp p = 0.0000 (p = 0.3514)
df - kapkoef & rentkoef p = 0.0000 (p = 0.6820)

The following edges may be removed from all identified models

Evidence		edge
Excl.	Incl.	

The following edges may only be removed (-) for a subset of models

Evidence			Model no.					
Excl.	Incl.	edge	1	2	3	4	5	6
0.066	0.002 *	ab - beinkak & beinkeqm	+	+	+	-	-	+
0.128	0.000 *	ae - beinkak & rdiffp	-	-	+	+	+	+
0.286	0.000 *	af - beinkak & rentkoef	+	+	-	+	+	-
0.094	0.000 *	be - beinkeqm & rdiffp	-	-	-	+	+	+
0.194	0.002 *	bf - beinkeqm & rentkoef	-	-	+	-	-	+
0.080	0.004 *	ce - beinkp & rdiffp	+	+	+	-	+	-
0.351	0.000 *	de - kapkoef & rdiffp	+	-	-	+	-	+
0.682	0.000 *	df - kapkoef & rentkoef	-	+	+	-	+	+
0.058	0.092	bc - beinkeqm & beinkp	-	-	-	+	+	+

b) Rückwärts-Elimination 1992

Model search finished.

bc: beinkeqm & beinkp eliminated
be: beinkeqm & rdiffp eliminated
bf: beinkeqm & rentkoef eliminated

a) EH-Algorithmus 1993

Step 1: Tests for conditional independence
The minimal model will be defined by removal of the following edges:

ab - beinkak & beinkeqm p = 0.0920
ac - beinkak & beinkp p = 0.1340
ae - beinkak & rdiffp p = 0.0517
af - beinkak & rentkoef p = 0.1240
bc - beinkeqm & beinkp p = 0.3000
be - beinkeqm & rdiffp p = 0.1200
bf - beinkeqm & rentkoef p = 0.0980
ce - beinkp & rdiffp p = 0.4300
de - kapkoef & rdiffp p = 0.2400
df - kapkoef & rentkoef p = 0.4420

Step 2: Examination of the adequacy of the minimal model
Evidence found in the following cases:

ac - beinkak & beinkp p = 0.0001 (p = 0.1340)
ae - beinkak & rdiffp p = 0.0000 (p = 0.0517)
af - beinkak & rentkoef p = 0.0000 (p = 0.1240)
bc - beinkeqm & beinkp p = 0.0280 (p = 0.3000)
be - beinkeqm & rdiffp p = 0.0006 (p = 0.1200)
bf - beinkeqm & rentkoef p = 0.0002 (p = 0.0980)
ce - beinkp & rdiffp p = 0.0142 (p = 0.4300)
de - kapkoef & rdiffp p = 0.0000 (p = 0.2400)
df - kapkoef & rentkoef p = 0.0000 (p = 0.4420)

The following edges may be removed from all identified models

Evidence		edge
Excl.	Incl.	
0.300	0.028	bc - beinkeqm & beinkp

The following edges may only be removed (-) for a subset of models

Evidence			Model no.								
Excl.	Incl.	edge	1	2	3	4	5	6	7	8	9
0.134	0.000 *	ac - beinkak & beinkp	-	-	-	-	-	+	+	-	+
0.052	0.000 *	ae - beinkak & rdiffp	+	+	+	+	+	+	+	+	-
0.124	0.000 *	af - beinkak & rentkoef	+	+	+	+	+	-	-	+	+
0.120	0.001 *	be - beinkeqm & rdiffp	+	-	-	-	-	-	-	+	+
0.098	0.000 *	bf - beinkeqm & rentkoef	-	+	+	-	-	+	+	-	-
0.430	0.014 *	ce - beinkp & rdiffp	-	-	-	-	-	-	-	+	+
0.240	0.000 *	de - kapkoef & rdiffp	+	+	-	+	-	+	-	-	+
0.442	0.000 *	df - kapkoef & rentkoef	-	-	+	-	+	-	+	+	-
0.092	0.092	ab - beinkak & beinkeqm	-	-	-	+	+	+	+	-	+

b) Rückwärts-Elimination 1993

Model search finished.

ab: beinkak & beinkeqm eliminated
ac: beinkak & beinkp eliminated
ad: beinkak & kapkoef eliminated

a) EH-Algorithmus 1994

Step 1: Tests for conditional independence
The minimal model will be defined by removal of the following edges:

ab - beinkak & beinkeqm p = 0.0656
ad - beinkak & kapkoef p = 0.1380
bc - beinkeqm & beinkp p = 0.4700
be - beinkeqm & rdiffp p = 0.1660
bf - beinkeqm & rentkoef p = 0.3160
cf - beinkp & rentkoef p = 0.2280
df - kapkoef & rentkoef p = 0.4020

Step 2: Examination of the adequacy of the minimal model
Evidence found in the following cases:

ad - beinkak & kapkoef p = 0.0016 (p = 0.1380)
be - beinkeqm & rdiffp p = 0.0400 (p = 0.1660)
bf - beinkeqm & rentkoef p = 0.0120 (p = 0.3160)
cf - beinkp & rentkoef p = 0.0441 (p = 0.2280)
df - kapkoef & rentkoef p = 0.0432 (p = 0.4020)

The following edges may be removed from all identified models

Evidence		edge
Excl.	Incl.	
0.470	0.241	bc - beinkeqm & beinkp

The following edges may only be removed (-) for a subset of models

Evidence			Model no.					
Excl.	Incl.	edge	1	2	3	4	5	6
0.138	0.002 *	ad - beinkak & kapkoef	+	-	-	-	+	+
0.166	0.040 *	be - beinkeqm & rdiffp	-	-	-	-	+	+
0.316	0.012 *	bf - beinkeqm & rentkoef	+	+	-	-	-	-
0.228	0.044 *	cf - beinkp & rentkoef	-	-	+	-	+	-
0.402	0.043 *	df - kapkoef & rentkoef	-	-	-	+	-	+
0.066	0.077	ab - beinkak & beinkeqm	-	+	+	+	-	-

b) Rückwärts-Elimination 1994

Model search finished.

ab: beinkak & beinkeqm eliminated
ac: beinkak & beinkp eliminated
ad: beinkak & kapkoef eliminated

```
df:  kapkoef & rentkoef   eliminated

Deviance  df   p       original df  model
326.2     311  0.2651  3600         (acde)(acef)(abd)
Degrees of freedom have been adjusted
```

```
be: beinkeqm &   rdiffp   eliminated
bf: beinkeqm & rentkoef   eliminated
ce:   beinkp &   rdiffp   eliminated
df:  kapkoef & rentkoef   eliminated

Deviance  df    p       original df  model
525.8     1282  1.0000  3951         (aef)(bcd)(cf)(de)
Degrees of freedom have been adjusted
```

```
ae:  beinkak &   rdiffp   eliminated
be: beinkeqm &   rdiffp   eliminated
bf: beinkeqm & rentkoef   eliminated
df:  kapkoef & rentkoef   eliminated

Deviance  df  p       original df  model
454.1     824 1.0000  924          (bcd)(cde)(cef)(af)
Degrees of freedom have been adjusted
```

Anhang Teil III

Umsetzung ausgewählter Methoden in Genstat Codes

Die Abschnitte, die mit einem Kommentar (Text in Anführungszeichen) in Großbuchstaben beginnen, erfordern eine Eingabe, zum Beispiel die Quelle der Datendatei. Rechenschritte, die keine weiter Aktion des Nutzers benötigen, sind durch einen Kommentar mit Kleinbuchstaben gekennzeichnet. Die Codes werden nach ihrem Auftauchen im Text in Kapitel 2 aufgelistet und zwar in der folgenden Reihenfolge:

Thema

1. Bipolare Korrespondenzanalyse
2. Nichtlineare Biplots
3. Hauptkomponenten Residuen
4. Velicers partielle Korrelations-Prozedur
5. Kreuzvalidierung
6. Hauptkomponenten-Biplots und CUSUM-Diagram
7. Gruppenanalysemodell und Gamma-q-q-Plots
8. Stabilitätsprüfung
9. Genstat-Menus zur Ergänzung der formalen Begriffsanalyse

1 Bipolare Korrespondenzanalyse

```
"Bipolare CA"


"OPENING THE DATASET"
open 'c:\\personal\\cyclamen\\data\\gs_cy4.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b


"POINTER VALUES AND MAXIMUM ON THE RATING SCALE"

"insert variate identifiers to form pointer"
pointer[values=s_ges_44,s_kno_44,s_wur_44,s_gil_44,s_wel_44,s_kra_44]data


"INSERT MAXIMUM AND MINIMUM VALUE ON THE RATING SCALE"
scalar[value=9]maxval
scalar[value=1]minval


"preliminary calculations"
calc data[]=data[]-minval
calc maxval=maxval-minval
calc nvars=nvalues(data)
calc nvals=nvalues(data[1])
calc nnvalsm1=(nvals-1)*-1
calc nvarsp1=nvars+1
calc dp_nvars=(2*nvars)
variate[nvalues=nvals;values=1...nvals]unit
text[nvalues=nvals]Betrieb
ftext unit;Betrieb


"LABELS FOR POSITIVE AND NEGATIVE POLS OF THE RATING SCALE"

"LABELS FOR THE UNITS (1) AND FOR THE VARIABLES"
text[nvalues=nvarsp1;values='Betrieb','Gesamt','Knospen','Wurzeln','Vergilbung','Welke',\
'Krankheiten']ratings


"DESCRIPTION FOR THE POLS"
text[nvalues=2;values='+','-']Pole


"LABELS FOR THE VARIABLES IN THE DOUBLED MATRIX"
text[nvalues=dp_nvars;values='ges+','kno+','wur+','gilb+','welke+','krank+',\
'ges-','kno-','wur-','gilb-','welke-','krank-']varname1;extra='Variable'


"+/- LABELS FOR THE VARIABLES IN THE DOUBLED MATRIX"
text[nvalues=nvars;values='Gesamt+/-','Knospen+/-','Wurzeln+/-',\
'Vergilbung+/-','Welke+/-','Krankheiten+/-']varname2;extra='Variable'


"SELECTION CRITERIA - 1 YES, 2 NO"
scalar[value=1]double


"double or simple matrix"

if double==1

"doubling tha data matrix"

variate[nvalues=nvals]maximum[1...nvars]
equate maxval;maximum

for ind=1...nvars
calc negdata[ind]=maximum[ind]-data[ind]
endfor

pointer[values=data[],negdata[]]combdata

variate[nvalues=dp_nvars;values=1...dp_nvars]colno
matrix[r=unit;c=colno]dop_mat
equate[oldf=!((1,#nnvalsm1)#dp_nvars,-1)]combdata;dop_mat

else

calc dp_nvars=nvars

variate[nvalues=dp_nvars;values=1...dp_nvars]colno
```

```
matrix[r=unit;c=colno]dop_mat
equate[oldf=!((1,#nnvalsm1)#dp_nvars,-1)]data;dop_mat

endif


"correspondence analysis"

corresp dop_mat;rowscores=rows;colscores=cols;roots=roots;rowinertia=rowin;colinertia=colin
calc rowscore[1,2]=rows$[*;1,2]
calc colscore[1,2]=cols$[*;1,2]


"calculations"

variate[nvalues=dp_nvars;values=(1...nvars)2]sorter
variate[nvalues=dp_nvars]cop1,cop2
equate colscore[1];cop1
equate colscore[2];cop2

for ind=1...nvars
variate[nvalues=2]colum1[ind],colum2[ind]
subset[condition=sorter.eq.ind]cop1;colum1[ind]
subset[condition=sorter.eq.ind]cop2;colum2[ind]
endfor


"producing the graph"

frame 1;xlower=0;xupper=0.7;ylower=0.3;yupper=1
frame 2;xlower=0.7;xupper=1;ylower=0;yupper=0.9
frame 3;xlower=0.2;xupper=1;ylower=0;yupper=0.3

pen 1;labels=Betrieb;linestyle=1
pen 2...nvarsp1;m=line;linestyle=1;symbol=2;labels=Pole
axes[equal=scale]1;style=box;xtitle='first dimension';ytitle='second dimension'

dgraph[keydescription='Objects and Variables';window=1;keywindow=3;\
title='Correspondence Analysis of Bipolar Data']\
rowscore[2],colum2[];rowscore[1],colum1[];description=ratings


"for output"

calc rootsumt=sum(roots)
scalar ro12[1...2]
equate roots;ro12
calc varex1=(ro12[1]/rootsumt)*100
calc varex2=(ro12[2]/rootsumt)*100
calc nvalro=nvalues(roots)

scalar rox[1...nvalro]
equate roots;rox


"absolute contributions"

calc colin12[1...2]=colin$[*;1,2]
calc colabcon[1]=colin12[1]/ro12[1]
calc colabcon[2]=colin12[2]/ro12[2]

calc rowin12[1...2]=rowin$[*;1,2]
calc rowabcon[1]=rowin12[1]/ro12[1]
calc rowabcon[2]=rowin12[2]/ro12[2]

calc rowcon1[1...nvalro]=(rowin$[*;1...nvalro])/rox[1...nvalro]
calc colcon1[1...nvalro]=(colin$[*;1...nvalro])/rox[1...nvalro]

calc rowcon2[1...nvalro]=(rowin$[*;1...nvalro])
calc colcon2[1...nvalro]=(colin$[*;1...nvalro])

factor[nvalues=nvalro;levels=nvalro;values=1...nvalro]PrinAxis
factor[nvalues=dp_nvars;levels=dp_nvars;values=1...dp_nvars]colvars
factor[nvalues=nvals;levels=nvals;values=1...nvals]rowvars
factor[nvalues=2;levels=2;values=1,2]Prinaxis
factor[nvalues=nvars;levels=nvars;values=1...nvars]colvars2

pen 1...20;labels=*;brush=16
axes[equal=no] 1;xmarks=*;ymarks=*;xtitle='principal component';ytitle=*

table[class=PrinAxis,rowvars]row1tab
equate rowcon2;row1tab
dbarchart[title='CUSUM Diagramm of Unit Contributions';window=1;keywindow=2;\
append=yes;keydescription='Units']row1tab
```

```
table[class=Prinaxis,rowvars]rowabtab
equate rowabcon;rowabtab
dbarchart[title='Absolute Unit Contributions';window=1;keywindow=2;\
append=yes;keydescription='Units']rowabtab

calc Unit_CON[1,2]=rowabcon[]*100

table[class=PrinAxis,colvars]col1tab
equate colcon2;col1tab
dbarchart[title='CUSUM diagramm of variable contributions';window=1;keywindow=2;\
append=yes;\
keydescription='Variables']col1tab;description=varname1

table[class=Prinaxis,colvars]colabtab
equate colabcon;colabtab
dbarchart[window=1;keywindow=2;\
title='Absolute Variable contributions';\
append=yes;keydescription='Variables']colabtab;description=varname1

calc Var_CON[1,2]=colabcon[]*100

variate[nvalues=dp_nvars]varcon1a,varcon2a
equate colabcon[1];varcon1a
equate colabcon[2];varcon2a

calc varcon1b=circulate(varcon1a;nvars)
calc varcon2b=circulate(varcon2a;nvars)
variate[nvalues=dp_nvars;values=1...dp_nvars]colcase
subset[condition=colcase<=nvars]varcon1a,varcon1b,varcon2a,varcon2b;vc1a,vc1b,vc2a,vc2b

calc vc1=vc1a+vc1b
calc vc2=vc2a+vc2b
pointer[values=vc1,vc2]vcpoint

pen 1...dp_nvars;colour=2...dp_nvars
table[class=Prinaxis,colvars2]col2tab
equate vcpoint;col2tab
dbarchart[title='Absolute Variable Contributions (accumulated)';window=1;keywindow=2;\
append=yes;keydescription='Variables']col2tab;description=varname2

variate[nvalues=nvals]rowinvar[1...nvalro]
equate rowcon2;rowinvar
variate[nvalues=dp_nvars]colinvar[1...nvalro]
equate colcon2;colinvar


"relative contributions"

calc inj=vsum(colinvar)
calc ink=vsum(rowinvar)

calc Var_COR[1...nvalro]=colinvar[]/inj
calc Unit_COR[1...nvalro]=rowinvar[]/ink

calc Var_QLT=Var_COR[1]+Var_COR[2]
calc Unit_QLT=Unit_COR[1]+Unit_COR[2]


"measures of polarisation"

calc zbarj[1...dp_nvars]=mean(combdata[])
calc kbarj[1...dp_nvars]=zbarj[]/maxval
calc polmj[1...dp_nvars]=1/(kbarj[]*(1-kbarj[]))

calc kij[1...dp_nvars]=combdata[]/maxval
calc m1kij[1...dp_nvars]=1-kij[]
calc inter1[1...dp_nvars]=kij[]*m1kij[]
calc inter2[1...dp_nvars]=sum(inter1[])
calc polei[1...dp_nvars]=nvals/inter2[]
calc polr[1...dp_nvars]=polmj[]/polei[]

variate[nvalues=nvars]polar1;extra='average'
variate[nvalues=nvars]polar2;extra='individual'
variate[nvalues=nvars]polar3;extra='relative'
equate polmj;polar1
equate polei;polar2
equate polr;polar3


"printed output"

print roots
print[iprint=*]'percentage of inertia explained through first principal axis',varex1
print[iprint=*]'percentage of inertia explained through second principal axis',varex2
```

```
print  'absolue unit contribution to the first two dimensions in %'
print[iprint=extra]Betrieb, Unit_CON[1],Unit_CON[2]

print  'absolue variable contribution to the first two dimensions in %'
print[iprint=extra]varname1, Var_CON[1],Var_CON[2]

print 'relative contributions and quality of the two-dimensional display of the units'
print[iprint=extra] Betrieb,Unit_QLT,Unit_COR[1],Unit_COR[2]

print 'relative contributions and quality of the two-dimensional display of the variables'
print[iprint=extra] varname1,Var_QLT,Var_COR[1],Var_COR[2]

print 'polarization'
print[iprint=extra]varname2,polar1,polar2,polar3
```

2 Nichtlineare Biplots

```
"Nichtlineare Biplots"


"OPENING THE DATASET"

open 'c:\\personal\\cyclamen\\data\\gs_cy4.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b

"DEFINITION OF DATA POINTER"
pointer[values=s_ges_44,s_kno_44,s_wur_44,s_gil_44,s_wel_44,s_kra_44]data

"DISTANCE"
text[value='city']test

"preliminaries 1"
calc nvars=nvalues(data)
calc nvals=nvalues(data[1])

"UNIT NAMES"
ftext betrieb;unitname

"VARIABLE NAMES AND UNITS DESCRIPTION"
text[values='Betrieb','Gesamt','Knospen','Wurzeln','Vergilbung','Welke',\
'Krankheiten']varname

"NUMBER OF BIPLOT TRAJECTORIES TO LOOK AT"
scalar[value=6]noofvars

"STEPSIZE"
variate[nvalues=nvars;values=1,1,1,1,1,1]varsteps[1...nvars]

"MINIMUM VALUE"
variate[nvalues=nvars;values=1,1,1,1,1,1]varmins[1...nvars]

"MAXIMUM NUMBER OF MARKERS"
scalar[value=1001]maxmarks


"preliminaries 2 - finding the marker labels"

calc noofvap1=noofvars+1
calc nodesc=-1*(nvars-noofvars)
text[nvalues=noofvap1]varsname
equate[oldf=!(noofvars,#nodesc)]varname;varsname

scalar stepsize[1...nvars]
scalar minvalue[1...nvars]

equate varsteps;stepsize
equate varmins;minvalue

calc maxmam1=maxmarks-1
variate[nvalues=maxmarks;values=0...#maxmam1]case[1...nvars]
calc case[1...nvars]=case[]*stepsize[]+minvalue[]
calc vals[1...nvars]=case[]

calc varmin[1...nvars]=min(data[1...nvars])
calc varmax[1...nvars]=max(data[1...nvars])
calc range[1...nvars]=varmax[1...nvars]-varmin[1...nvars]

matrix[r=nvals;c=nvars]m
calc n_m1=-1*(nvals-1)
equate[oldf=!((1,#n_m1)#nvars,-1)]data;m

for ind=1...nvars
subset[condition=vals[ind]<=varmin[ind]]case[ind];lowcase[ind]
endfor

for ind=1...nvars
subset[condition=vals[ind]>=varmax[ind]]case[ind];upcase[ind]
endfor

calc nonvals[1...nvars]=nvalues(lowcase[1...nvars])\
+nvalues(upcase[1...nvars])-2
calc Nvals[1...nvars]=maxmam1-nonvals[1...nvars]

for ind=1...nvars
variate[nvalues=Nvals[ind]]marker[ind]
endfor
```

```
calc max_lowc[1...nvars]=max(lowcase[1...nvars])
calc min_upc[1...nvars]=min(upcase[1...nvars])

for ind=1...nvars
subset[condition=case[ind]>=max_lowc[ind] .and. case[ind]<=min_upc[ind]]\
vals[ind];marker[ind]
endfor


"centre data"

scal xbar[1...nvars]
calc cdata[1...nvars] = data[] - (xbar[] = mean(data[]))
calc max[1...nvars] = max(cdata[])
calc min[1...nvars] = min(cdata[])
calc rng[1...nvars] = max[] - min[]
calc maxrng = vmax(rng)


"calculating centred marker"

calc cmarker[1...nvars]=marker[]-xbar[]


"adding the means for the point of concurrency"

variate[nvalues=1;value=0]zero[1...nvars]
variate[nvalues=1]databar[1...nvars]
equate xbar;databar

for ind=1...nvars
append marker[ind],databar[ind]
append cmarker[ind],zero[ind]
sort marker[ind]
sort cmarker[ind]
endfor


"calculating number of markers"

calc nvarsm1=nvars-1
calc marval[1...nvars]=nvalues(marker[])
calc sum_mark=vsum(marval)
calc dif_mark[1...nvars]=sum_mark-marval[]


"pco"

fsim[similarity=sim]cdata[];test=#nvars(#test);range=maxrng
pco[p=roots] sim;dist=dist;lrv=l;centroid=c
calc psco[1,2]=l[1]$[*;1,2]

axes[equal=scale]1;style=box;xorigin=*;yorigin=*;xtitle='first dimension';\
ytitle='second dimension'
pen 1;labels=unitname
pen 2;linestyle=1

dgraph[keywindow=0;title='Principal Coordinates Analysis, centred data']\
psco[2];psco[1]

dmst[dimensions=2,1;title='PCO with MST, centred data']\
coordinates=l[1];similarity=sim;\
symbols=unitname


"calculating pseudo variables"

variate[nvalues=sum_mark]pseudo[1...nvars]

for ind=1...nvars
variate[nvalues=dif_mark[ind]]zeros[ind]
calc zeros[ind]=mean(cdata[ind])
equate !P(cmarker[ind],zeros[ind]);pseudo[ind]
endfor

variate[nvalues=nvars]oneval
equate marval;oneval
calc acc_one=cum(oneval)

scalar shifter[1...nvars]
equate acc_one;shifter

for ind=1...nvars
calc basis[ind]=shifter[ind]-marval[ind]
calc pseudo[ind]=circulate(pseudo[ind];basis[ind])
```

```
endfor


"forming new (pseudo and old data) dataset"

calc newno=sum_mark+nvals
variate[nvalues=newno]newdata[1...nvars]
for ind=1...nvars
equate !P(cdata[ind],pseudo[ind]);newdata[ind]
endfor


"calculating new distances"

fsim[similarity=newsim]newdata[];test=#nvars(#test);range=maxrng
calc newdist=1-newsim


"forming the trajectories"

symmetricmatrix[r=nvals]oldsim
calc oldsim=newsim$[!(1...nvals)]

for ind=1...nvars
calc first[ind]=nvals+shifter[ind]-marval[ind]+1
calc last[ind]=nvals+shifter[ind]
matrix[r=marval[ind];c=nvals]ps_sim[ind]
calc ps_sim[ind]=newdist$[!(first[ind]...last[ind]);!(1...nvals)]
endfor

lrv[r=nvals;c=2]l2
pco sim;dist=dist;lrv=l2;centroid=c

for ind =1...nvars
addpoints ps_sim[ind];lrv=l2;centroid=c;coordinates=trajemat[ind]
endfor

for ind=1...nvars
calc tra_vals[ind]=nvalues(marker[ind])
variate[nvalues=tra_vals[ind]]trajector[1,2][ind]
calc trajector[1,2][ind]=trajemat[ind]$[*;1,2]
endfor


"making the graph"

for ind=1...nvars
ftext marker[ind];variable[ind];decimals=0
endfor

pen 2...noofvap1;method=line;linestyle=1;symbol=0;\
labels=variable[];size=0.75;join=given

dgraph[keywindow=0;title='Interpolative Nonlinear Biplots']\
psco[2],trajector[2][];\
psco[1],trajector[1][]


"streching the trajectories"

subset[condition=marker[1].eq.xbar[1]]trajector[1][1];correct[1]
subset[condition=marker[1].eq.xbar[1]]trajector[2][1];correct[2]
scalar cct[1,2]
equate correct;cct

calc tra2[1][1...nvars]=((trajector[1][]-cct[1])*nvars)+cct[1]
calc tra2[2][1...nvars]=((trajector[2][]-cct[2])*nvars)+cct[2]

frame 1;xlower=0;xupper=0.7;ylower=0.3;yupper=1
frame 2;xlower=0.2;xupper=1;ylower=0;yupper=0.3

pen 1;method=point;symbol=1;labels=unitname;size=1
pen 2...noofvap1;method=line;symbol=0;labels=variable[];\
size=0.75;join=given;linestyle=1

dgraph[keywindow=2;keydescription='Objects and Variables';\
title='Interpolative Nonlinear Biplot (streched trajectories)']\
psco[2],tra2[2][];\
psco[1],tra2[1][];description=varsname
```

3 Hauptkomponenten Residuen

```
"PCA Residuen"


"NAME AND LOCATION OF DATA FILE"
open 'c:\\personal\\cyclamen\\data\\gs_cy3_3.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b

"DEFINITION OF DATA AND PCA MODEL"
pointer[values=n23,n29,n41,k23,k29,k41,ph23,ph29,ph41,salz23,salz29,salz41]data
text[nvalues=1;value='yes']standard
scalar[value=2]model
scalar[value=0.05]alpha

calc nvals=nvalues(data[1])
calc nvars=nvalues(data)
variate[nvalues=nvals;values=1,2,3,4,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]case


"standardisation"

if standard .eqs. 'yes'

calculate variance[1...nvars]=var(data[])
calculate stddev[1...nvars]=sqrt(variance[])

for ind=1...nvars
if stddev[ind]>0
calculate zdata[ind]=(data[ind]-mean(data[ind]))/stddev[ind]
else
calculate zdata[ind]=('missing')
endif
endfor

calc ydata[1...nvars]=data[]
calc data[]=zdata[]

elsif standard .eqs. 'no'
calc ydata[1...nvars]=data[]

endif

matrix[r=nvals;c=1]rest
pcp[nroots=model;m=var]data;lrv=l;scores=score;residual=rest
calc sqrtroot=sqrt(l[2])
scalar sval[1...model]
equate sqrtroot;sval
calc yscore[1...model]=score$[*;1...model]
calc yscore[1...model]=yscore[1...model]/sval[1...model]

calc  ysquare[1...model]=yscore[]**2
variate[nvalues=nvals]y[1...model]
equate ysquare;y

calc Tsquare=vsums(y)
calc Q=rest**2


"graph of residuals"

text[nvalues=nvals]unitname
ftext case;unitname
frame 1;xlower=0.1;xupper=0.9;ylower=0.1;yupper=0.9
axes 1;xtitle='Index';ytitle='Residual';xmarks=10
pen 1;labels=unitname
dgraph[title='PCA Residuals';keywindow=0]Q;case

dgraph[title='PCA Residuals';keywindow=0]Q;case


"citical values"

pcp[m=var]data;lrv=l2
scalar lj[1...nvars]
equate l2[2];lj

calc j=model+1
pointer[values=lj[j...nvars]]p1
calc theta1=vsums(p1)

calc p2[j...nvars]=p1[]**2
calc theta2=vsums(p2)
```

```
calc p3[j...nvars]=p1[]*p2[]
calc theta3=vsums(p3)

calc ho=1-(2*theta1*theta3)/(3*theta2**2)
calc calpha=ednormal(alpha)

if ho>0
calc calpha=calpha*-1
elsif ho<=0
calc calpha=calpha
endif

calc qa1=(calpha*sqrt(2*theta2*ho**2))/theta1
calc qa2=(theta2*ho*(ho-1))/theta1**2
calc qa3=(qa1+qa2+1)**(1/ho)
calc qa4=theta1*qa3

calc qalpha=qa4

calc malpha=1-alpha
calc n_p=nvals-model
calc fal=fed(malpha;model;n_p)
calc critT=(model*(nvals-1)/(nvals-model))*fal


"printed output"

print 'Q-values (PCA residuals)'
print unitname,Q

print[iprint=*] 'Critical value at alpha',alpha,'for Q'
print[iprint=*]qalpha

print 'T2-values'
print unitname,Tsquare

print[iprint=*]'Critical value at alpha',alpha,'for T2'
print[iprint=*]critT
```

4 Velicers partielle Korrelations-Prozedur

```
"Velicers partielle Korrelations-Prozedur"


"OPENING THE DATA SET"

open 'c:\\personal\\cyclamen\\data\\gs_cy3_1.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b

"DEFINITIONS"

pointer[values=salz23,salz29,salz41,ph23,ph29,ph41,n23,n29,n41,k23,k29,k41]data
text[nvalues=1;values='yes']standard


"preliminaries"

if standard .eqs. 'yes'
calc nvars=nvalues(data)
calc v[1...nvars]=var(data[1...nvars])
calc s[1...nvars]=sqrt(v[1...nvars])
calc data[1...nvars]=(data[1...nvars]-mean(data[1...nvars]))/s[1...nvars]
elsif standard .eqs. 'no'
calc nvars=nvalues(data)
endif
calc nvals=nvalues(data[1])


"calculation of the covariance matrix"

sspm[t=data[1...nvars]]ssp;ssp=sp;nunits=nu
fsspm ssp
calc df=nu-1
calc covmat=1/df*sp


"calculation of square roots of eigenvalues"

pcp[m=corr;p=score]data;lrv=l

scalar rootvar[1...nvars]
equate l['roots'];rootvar
for ind=1...nvars
if rootvar[ind].eq.0
calc svals=0
else
calc rootvar[ind]=sqrt(rootvar[ind])
endif
endfor

scalar sinval[1...nvars]
equate rootvar;sinval


"calculation of v-vectors"

matrix[r=nvars;c=1]m[1...nvars]
calc mt=t(l['vectors'])
equate mt;m
matrix[r=nvars;c=1]vvector[1...nvars]
calc vvector[]=sinval[]*m[]


"calculation of the matrices sq"

matrix[r=nvars;c=nvars]vv[1...nvars]
calc vv[]=rtproduct(vvector[];vvector[])
scalar[value=0]null
matrix[r=nvars;c=nvars]sv[0...nvars]
equate null;sv[0]
calc nvarsm1=nvars-1
calc sv[1...nvars]=sv[0...nvarsm1]+vv[1...nvars]
matrix[r=nvars;c=nvars]sq[0...nvarsm1]
calc sq[]=sv[nvars]-sv[0...nvarsm1]


"calculation of residual matrices rq"

calc nvarsneg=-1*nvars
diagonal[r=nvars]dsq[0...nvarsm1]
equate[oldf=!(1,nvarsneg)]sq[];dsq[]
```

```
diagonal[r=nvars]dsq_p12[0...nvarsm1]
calc dsq_p12[]=dsq[]**(-1/2)
matrix[r=nvars;c=nvars]rq[0...nvarsm1]
calc rq[]=product(dsq_p12[];sq[])
calc rq[]=product(rq[];dsq_p12[])


"calculation of fq values"

scalar riJ[0...nvarsm1]
calc riJ[]=sum(rq[]**2)-nvars
scalar fq[0...nvarsm1]
calc fq[]=riJ[]/(nvars*nvarsm1)
variate[nvalues=nvars;values=0...nvarsm1]princomp;decimals=0
variate[nvalues=nvars;values=fq[]]fqvalues


"output and presentatation of results"

frame window=1;ylower=0;yupper=1;xlower=0;xupper=1
axes window=1;xtitle='principal components';ytitle='fq values';xmarks=1

dgraph[t='Velicers fq-values';keyw=0]fqvalues;princomp

print 'Results of the partial correlation procedure (Velicer, 1976)'
print 'Principal components, fq-values'
print[clprint=*]princomp,fqvalues
```

5 Kreuzvalidierung

```
"Kreuzvalidation"


"NAME AND LOCATION OF DATA FILE"

open 'c:\\personal\\cyclamen\\data\\gs_cy3_3.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b

"DEFINITION OF DATA POINTER"
pointer[values=n23,k23,ph23,salz23,n29,n41,k29,k41,ph29,ph41,\
salz29,salz41]data

"STANDARDISATION"
text[nvalues=1;values='yes']standard


"preliminaries"

calc nvals=nvalues(data[1])
calc nvars=nvalues(data)
calc n=nvals*nvars
calc varn=n*nvars
calc nvarsm1=nvars-1
calc nvars2=nvarsm1*nvars
calc circ=nvals*nvarsm1
calc nvalsm1=nvals-1
calc nmnvars=n-nvars
calc n2=n*nvarsm1


"standardisation of data"

if standard .eqs. 'yes'

calculate variance[1...nvars]=var(data[])
calculate stddev[1...nvars]=sqrt(variance[])

for ind=1...nvars
if stddev[ind]>0
calculate zdata[ind]=(data[ind]-mean(data[ind]))/stddev[ind]
else
calculate zdata[ind]=('missing')
endif
endfor

calc ydata[1...nvars]=data[]
calc data[]=zdata[]

elsif standard .eqs. 'no'
calc ydata[1...nvars]=data[]

endif


"column Reduction"
"equate pointer to variate and circulate"

variate[nvalues=n]datavar
equate data;datavar
variate[nvalues=n]datvar_l[0...nvars]
calc datvar_l[0]=datavar
calc datvar_l[1...nvars]=circulate(datvar_l[0...nvarsm1];circ)

"create sorting variates"
variate[nvalues=n;values=1...n]case1
variate[nvalues=n;values=1...n]case2[0...nvars]
calc case2[1...nvars]=circulate(case2[0...nvarsm1];circ)

"create column reduced data sets"
calc limit=nvals*nvars-nvals+1
subset[condition=case1<limit]datvar_l[],case2[]

"sort column reduced data sets"
for i=1...nvars
sort[index=case2[i]]case2[i],datvar_l[i]
endfor

"create column reduced data matrices"
matrix[r=nvals;c=nvarsm1]credmat[1...nvars]
matrix[r=nvarsm1;c=nvals]tcredmat[1...nvars]
```

```
equate datvar_l[1...nvars];tcredmat[1...nvars]
calc credmat[]=t(tcredmat[])

delete[redefine=yes;list=exclusive]data,nvals,nvars,n,varn,nvarsm1,\
nvars2,circ,nvalsm1,nmnvars,n2,credmat,datavar


"row Reduction"
"generate sorting factors"

factor[nvalues=n;levels=nvars]f1
factor[nvalues=n;levels=nvals]f2
generate f1,f2

"create row reduced data set"

variate[nvalues=nmnvars]redvar[1...nvals]
for i=1...nvals
subset[condition=f2.ne.i]datavar
calc redvar[i]=datavar
variate[nvalues=n]datavar
equate data;datavar
endfor

"create row reduced data matrices"

matrix[r=nvars;c=nvalsm1]trredmat[1...nvals]
matrix[r=nvalsm1;c=nvars]rredmat[1...nvals]
equate redvar;trredmat
calc rredmat[]=t(trredmat[])

delete[redefine=yes;list=exclusive]data,nvals,nvars,n,varn,nvarsm1,\
nvars2,circ,nvalsm1,nmnvars,n2,credmat,rredmat


"singular value decomposition of reduced data matrices"
"SVDs for all matrices"
"hat types equal column reduced matrices"
"bar types equal row reduced matrices"

diagonal[r=nvars]Shat[1...nvars]
diagonal[r=nvals]Sbar[1...nvals]
matrix[r=nvals;c=nvarsm1]Uhat[1...nvars]
matrix[r=nvalsm1;c=nvars]Vbar[1...nvals]
svd credmat[1...nvars];l=Uhat[1...nvars];s=Shat[1...nvars]
svd rredmat[1...nvals];s=Sbar[1...nvals];r=Vbar[1...nvals]

for ind=1...nvars
variate[nvalues=nvals]uhatv[ind][1...nvarsm1]
calc tUhat[ind]=t(Uhat[ind])
equate tUhat[ind];uhatv[ind]
scalar shatsc[ind][1...nvarsm1]
equate Shat[ind];shatsc[ind]
endfor

for ind=1...nvals
variate[nvalues=nvars]vbarv[ind][1...nvars]
calc tVbar[ind]=t(Vbar[ind])
equate tVbar[ind];vbarv[ind]
scalar sbarsc[ind][1...nvars]
equate Sbar[ind];sbarsc[ind]
endfor

delete[redefine=yes;list=exclusive]data,nvals,nvars,n,varn,nvarsm1,\
nvars2,circ,nvalsm1,nmnvars,n2,uhatv,shatsc,vbarv,sbarsc


"preliminary data manipulations to calculate xijhat"

for i=1...nvars
for j=1...nvarsm1
calc part1[i][j]=uhatv[i][j]*sqrt(shatsc[i][j])
endfor
endfor

for i=1...nvals
for j=1...nvars
calc part2[i][j]=vbarv[i][j]*sqrt(sbarsc[i][j])
endfor
endfor


variate[nvalues=varn]part2var
append[newvector=part2var]part2[][]
variate[nvalues=varn;values=(#nvars2(1),#nvars(2))#nvals]sorter1
```

```
variate[nvalues=n2]part2va
subset[condition=sorter1.eq.1]part2var;part2va

variate[nvalues=n2;values=(1...nvars2)#nvals]sorter2
for ind=1...nvars2
subset[condition=sorter2.eq.ind]part2va;part2new[ind]
endfor

variate[nvalues=n2]part1var
append[newvector=part1var]part1[][]
variate[nvalues=n2;values=#nvals(1...nvarsm1)#nvars]sorter3

for ind=1...nvarsm1
variate[nvalues=n]part1v[ind]
subset[condition=sorter3.eq.ind]part1var;part1v[ind]
endfor

variate[nvalues=nvals]part1new[1...nvars2]
equate part1v;part1new

variate[nvalues=nvals]uhsh[1...nvarsm1][1...nvars]
variate[nvalues=nvals]sbvb[1...nvarsm1][1...nvars]
calc uhsh[][]=part1new[]
calc sbvb[][]=part2new[]

delete[redefine=yes;list=exclusive]data,nvals,nvars,n,varn,nvarsm1,\
nvars2,circ,nvalsm1,nmnvars,n2,uhsh,sbvb


"calculation of xijhat"

calc xijhat[1...nvarsm1][1...nvars]=uhsh[][]*sbvb[][]


"control"

matrix[r=nvals;c=nvars]mats
matrix[r=nvars;c=nvals]tmats
equate data;tmats
calc mats=t(tmats)
svd mats;l=U;s=S;r=V


"parity check"

calc Uo[1...nvars]=U
calc Vo[1...nvals]=V
calc Shato[1...nvars]=S
calc Sbaro[1...nvals]=S

for ind=1...nvars
variate[nvalues=nvals]Uhatvo[ind][1...nvarsm1]
calc tUo[ind]=t(Uo[ind])
equate tUo[ind];Uhatvo[ind]
scalar Shatsco[ind][1...nvarsm1]
equate Shato[ind];Shatsco[ind]
endfor

for ind=1...nvals
variate[nvalues=nvars]Vbarvo[ind][1...nvars]
calc tVo[ind]=t(Vo[ind])
equate tVo[ind];Vbarvo[ind]
scalar Sbarsco[ind][1...nvars]
equate Sbaro[ind];Sbarsco[ind]
endfor

for i=1...nvars
for j=1...nvarsm1
calc part1o[i][j]=Uhatvo[i][j]*sqrt(Shatsco[i][j])
endfor
endfor

for i=1...nvals
for j=1...nvars
calc part2o[i][j]=Vbarvo[i][j]*sqrt(Sbarsco[i][j])
endfor
endfor

variate[nvalues=varn]p2ovar
append[newvector=p2ovar]part2o[][]
variate[nvalues=varn;values=(#nvars2(1),#nvars(2))#nvals]sorter1
variate[nvalues=n2]p2ova
subset[condition=sorter1.eq.1]p2ovar;p2ova

variate[nvalues=n2;values=(1...nvars2)#nvals]sorter2
```

```
for ind=1...nvars2
subset[condition=sorter2.eq.ind]p2ova;p2onew[ind]
endfor

variate[nvalues=n2]p1ovar
append[newvector=p1ovar]part1o[][]
variate[nvalues=n2;values=#nvals(1...nvarsm1)#nvars]sorter3

for ind=1...nvarsm1
variate[nvalues=n]p1ov[ind]
subset[condition=sorter3.eq.ind]p1ovar;p1ov[ind]
endfor

variate[nvalues=nvals]p1onew[1...nvars2]
equate p1ov;p1onew

variate[nvalues=nvals]uhsho[1...nvarsm1][1...nvars]
variate[nvalues=nvals]sbvbo[1...nvarsm1][1...nvars]
calc uhsho[][]=p1onew[]
calc sbvbo[][]=p2onew[]

calc xijhato[1...nvarsm1][1...nvars]=uhsho[][]*sbvbo[][]
calc xijhatn[1...nvarsm1][1...nvars]=xijhat[][]
calc parity1[1...nvarsm1][1...nvars]=xijhato[][]*xijhat[][]

for i=1...nvarsm1
for j=1...nvars
calc minpar[i][j]=min(parity1[i][j])
if minpar[i][j].lt.0
restrict xijhatn[i][j];condition=parity1[i][j].lt.0
calc xijhatn[i][j]=xijhatn[i][j]*-1
restrict xijhatn[i][j]
endif
endfor
endfor

equate xijhatn[];xijhat[]

delete[redefine=yes;list=exclusive]data,nvals,nvars,n,varn,nvarsm1,\
nvars2,circ,nvalsm1,nmnvars,n2,xijhat


"calculation of PRESS"

for ind=1...nvarsm1
variate[nvalues=n]Mhat[ind]
equate xijhat[ind];Mhat[ind]
endfor

variate[nvalues=n]xij
equate data;xij

variate[nvalues=n;values=#n(0)]M0
scalar[value=0]DM0
calc press0=(sum(xij**2))/n

for ind=1...nvarsm1
calc xijhatM[ind]=M0+Mhat[ind]
calc M0=xijhatM[ind]
calc press[ind]=(sum((xijhatM[ind]-xij)**2))/n
calc DM[ind]=nvals+nvars-2*ind
calc DM0=DM0+DM[ind]
calc DR[ind]=nvars*nvalsm1-DM0
calc W[ind]=((press0-press[ind])/DM[ind])/(press[ind]/DR[ind])
calc press0=press[ind]
endfor


"output"

variate[nvalues=nvarsm1]Wvar
equate W;Wvar
variate[nvalues=nvarsm1;values=1...nvarsm1]pc
variate[nvalues=nvarsm1;values=#nvarsm1(1)]lims
pen 1;s=2
pen 2;s=0;m=l
axes 1;xtitle='Principal Components';ytitle='W-values';xmarks=1
dgraph[keywindow=0;title='PRESS statistics']Wvar,lims;pc,pc

axes 1;xtitle='Principal Components';ytitle='W-values';xmarks=1
dgraph[keywindow=0;title='PRESS statistics']Wvar,lims;pc,pc

print 'pc','      PRESS values W'
print[iprint=*]pc,Wvar
```

6 Hauptkomponenten-Biplots und CUSUM-Diagram

```
"Hauptkomponenten-Biplots mit Interpolations- und Prediktionsmarkern und CUSUM-Diagram"


"NAME AND LOCATION OF DATA FILE"
open 'c:\\phd\\cyclamen\\data\\gs_cy3_3.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b


"DEFINITION OF DATA POINTER"
pointer[values=n23,k23,ph23,salz23,n29,k29,ph29,salz29,\
n41,k41,ph41,salz41]data

calc nvars=nvalues(data)

"UNIT NAMES"
text[values='1','2','3','4','6','7','8','9','10',\
'11','12','13','14','15','16','17','18','19','20']unitname

"VARIABLE NAMES AND UNITS DESCRIPTION"
text[values='N Woche 23','K Woche 23','pH Woche 23','Salz Woche 23',\
            'N Woche 29','K Woche 29','pH Woche 29','Salz Woche 29',\
            'N Woche 41','K Woche 41','pH Woche 41','Salz Woche 41',\
            'Betriebe']varname

"NUMBER OF BIPLOT AXES TO LOOK AT"
scalar[value=4]noofvars

"STEPSIZE"
variate[nvalues=nvars;values=#nvars(1)]varsteps[1...nvars]

"MINIMUM VALUE"
variate[nvalues=nvars;values=#nvars(-3)]varmins[1...nvars]

"MAXIMUM NUMBER OF MARKERS"
scalar[value=1001]maxmarks

"STANDARDISATION"
text[nvalues=1;values='yes']standard


"preliminaries"

calc noofvap1=noofvars+1
calc nodesc=-1*(nvars-noofvars)
text[nvalues=noofvap1]varsname
equate[oldf=!(noofvars,#nodesc)]varname;varsname

scalar stepsize[1...nvars]
scalar minvalue[1...nvars]

equate varsteps;stepsize
equate varmins;minvalue
calc ydata[1...nvars]=data[]


"standardisation"

if standard .eqs. 'yes'

calculate variance[1...nvars]=var(data[])
calculate stddev[1...nvars]=sqrt(variance[])

for ind=1...nvars
if stddev[ind]>0
calculate zdata[ind]=(data[ind]-mean(data[ind]))/stddev[ind]
else
calculate zdata[ind]=('missing')
endif
endfor

calc data[]=zdata[]

elsif standard .eqs. 'no'
calc data[]=data[]

endif


"marker 1"
```

```
calc maxmam1=maxmarks-1
variate[nvalues=maxmarks;values=0...#maxmam1]case[1...nvars]
calc case[1...nvars]=case[]*stepsize[]+minvalue[]
calc vals[1...nvars]=case[]

calc nvars=nvalues(data)
calc varmin[1...nvars]=min(data[1...nvars])
calc varmax[1...nvars]=max(data[1...nvars])
calc range[1...nvars]=varmax[1...nvars]-varmin[1...nvars]

calc n=nvalues(data[1])
matrix[r=n;c=nvars]m
calc n_m1=-1*(n-1)
equate[oldf=!((1,#n_m1)#nvars,-1)]data;m

for ind=1...nvars
subset[condition=vals[ind]<=varmin[ind]]case[ind];lowcase[ind]
endfor

for ind=1...nvars
subset[condition=vals[ind]>=varmax[ind]]case[ind];upcase[ind]
endfor

calc nonvals[1...nvars]=nvalues(lowcase[1...nvars])\
+nvalues(upcase[1...nvars])-2
calc nvals[1...nvars]=maxmam1-nonvals[1...nvars]

for ind=1...nvars
variate[nvalues=nvals[ind]]marker[ind]
endfor

calc max_lowc[1...nvars]=max(lowcase[1...nvars])
calc min_upc[1...nvars]=min(upcase[1...nvars])

for ind=1...nvars
subset[condition=case[ind]>=max_lowc[ind] .and. case[ind]<=min_upc[ind]]\
vals[ind];marker[ind]
endfor


"marker 2"

pcp[m=variance]data;lrv=lrv;scores=sc
variate[nvalues=n]units[1...2]
calc units[]=sc$[*;1,2]


"marker 3"

calc meandata[1...nvars]=mean(data[])
calc diff[1...nvars]=marker[]-meandata[]
calc absdiff[1...nvars]=abs(marker[]-meandata[])
calc mindiff[1...nvars]=min(absdiff[])
calc nmarks[1...nvars]=nvalues(marker[])

for ind=1...nvars
variate[nvalues=nmarks[ind];values=1...nmarks[ind]]mult[ind]
subset[condition=absdiff[ind].eq.mindiff[ind]]mult[ind];factor[ind]
variate[nvalues=nmarks[ind];values=#nmarks[ind](factor[ind])]difffac[ind]
calc mue[ind]=stepsize[ind]*(mult[ind]-difffac[ind])
subset[condition=mue[ind].eq.0]diff[ind];lambda[ind]
endfor

variate[nvalues=nvars]vrhovar[1...nvars]
calc nnvarsm1=-1*(nvars-1)
equate[oldf=!((1,#nnvarsm1)#nvars,-1)]lrv[1];vrhovar
matrix[r=nvars;c=1]vrho1
matrix[r=nvars;c=1]vrho2
equate vrhovar[1];vrho1
equate vrhovar[2];vrho2

variate[nvalues=nvars]lambvar
equate lambda;lambvar
calc p11=lambvar*vrho1
calc p12=lambvar*vrho2
scalar part11[1...nvars]
scalar part12[1...nvars]
equate p11;part11
equate p12;part12

scalar vrhok1[1...nvars]
scalar vrhok2[1...nvars]
equate vrho1;vrhok1
equate vrho2;vrhok2
```

```
for ind=1...nvars
calc part21[ind]=mue[ind]*vrhok1[ind]
calc part22[ind]=mue[ind]*vrhok2[ind]
calc xaxes[ind]=part11[ind]+part21[ind]
calc yaxes[ind]=part12[ind]+part22[ind]
endfor


"marker 4"

calc nvarsp1=noofvars+1

for ind=1...nvars
ftext marker[ind];text=tm[ind]
endfor

axes[equal=scale]1;style=box;\
xtitle='first component';ytitle='second component'
pen nvarsp1;labels=unitname;colour=1;m=point;symbol=1;size=1
pen 1...noofvars;m=line;linestyle=1;\
symbols=2;labels=tm[1...noofvars];size=0.75
frame 1;xlower=0;xupper=0.9;ylower=0.1;yupper=1
frame 2;xlower=0.9;xupper=1.3;ylower=0;yupper=0.9

dgraph[keywindow=2;keydescription='Objects and Units';\
title='Interpolation PCA Biplot']\
yaxes[1...noofvars],units[2];\
xaxes[1...noofvars],units[1];description=varsname


"marker 5"

calc meandata[1...nvars]=mean(data[])
calc diff[1...nvars]=marker[]-meandata[]
calc absdiff[1...nvars]=abs(marker[]-meandata[])
calc mindiff[1...nvars]=min(absdiff[])
calc nmarks[1...nvars]=nvalues(marker[])

for ind=1...nvars
variate[nvalues=nmarks[ind];values=1...nmarks[ind]]mult[ind]
subset[condition=absdiff[ind].eq.mindiff[ind]]mult[ind];factor[ind]
variate[nvalues=nmarks[ind];values=#nmarks[ind](factor[ind])]difffac[ind]
calc mue[ind]=stepsize[ind]*(mult[ind]-difffac[ind])
subset[condition=mue[ind].eq.0]diff[ind];lambda[ind]
endfor

variate[nvalues=nvars]vrhovar[1...nvars]
calc nnvarsm1=-1*(nvars-1)
equate[oldf=!((1,#nnvarsm1)#nvars,-1)]lrv[1];vrhovar
matrix[r=nvars;c=1]vrho1
matrix[r=nvars;c=1]vrho2
equate vrhovar[1];vrho1
equate vrhovar[2];vrho2

scalar vrhok1[1...nvars]
scalar vrhok2[1...nvars]
equate vrho1;vrhok1
equate vrho2;vrhok2

calc nvarsm1=nvars-1
variate[nvalues=nvars;values=1,#nvarsm1(0)]ej[1...nvars]
calc ej[2...nvars]=circulate(ej[1...nvarsm1])
matrix[r=1;c=nvars]ek[1...nvars]
equate ej;ek

matrix[r=nvars;c=2]vrho
equate[oldf=!((1,nnvarsm1)2,-1)]!P(vrho1,vrho2);vrho

scalar resultx1[1...nvars]
scalar resulty1[1...nvars]

for ind=1...nvars
calc divisor[ind]=ek[ind]*+vrho*+t(vrho)*+t(ek[ind])
calc resx1[ind]=vrhok1[ind]/divisor[ind]
calc resy1[ind]=vrhok2[ind]/divisor[ind]
endfor

equate resx1;resultx1
equate resy1;resulty1

variate[nvalues=nvars]resultx2,resulty2
equate resx1;resultx2
equate resy1;resulty2

variate[nvalues=nvars]lambvar
```

```
equate lambda;lambvar
calc pp11=lambvar*resultx2
calc pp12=lambvar*resulty2
scalar ppart11[1...nvars]
scalar ppart12[1...nvars]
equate pp11;ppart11
equate pp12;ppart12

for ind=1...nvars
calc ppart21[ind]=mue[ind]*resultx1[ind]
calc ppart22[ind]=mue[ind]*resulty1[ind]
calc pxaxes[ind]=ppart11[ind]+ppart21[ind]
calc pyaxes[ind]=ppart12[ind]+ppart22[ind]
endfor


"marker 6"

if standard .eqs. 'no'

calc nvarsp1=noofvars+1

for ind=1...nvars
ftext marker[ind];text=tm[ind]
endfor

axes[equal=scale]1;style=box;\
xtitle='first component';ytitle='second component'
pen nvarsp1;labels=unitname;colour=1;symbols=1;size=1;m=point
pen 1...noofvars;m=line;linestyle=1;\
symbols=2;labels=tm[1...noofvars];size=1

frame 1;xlower=0;xupper=0.8;ylower=0.2;yupper=1
frame 2;xlower=0.85;xupper=1.3;ylower=0;yupper=0.9

dgraph[keywindow=2;keydescription='Objects and Variables';\
title='Prediction PCA Biplot']\
pyaxes[1...noofvars],units[2];\
pxaxes[1...noofvars],units[1];description=varsname


"marker 6 ZI"

elsif standard .eqs. 'yes'

calc nvarsp1=noofvars+1

for ind=1...nvars
ftext marker[ind];text=tm[ind]
endfor

axes[equal=scale]1;style=box;\
xtitle='first component';ytitle='second component'
pen nvarsp1;labels=unitname;colour=1;symbols=1;size=1;m=point
pen 1...noofvars;m=line;linestyle=1;\
symbols=2;labels=tm[1...noofvars];size=1

frame 1;xlower=0;xupper=0.9;ylower=0.1;yupper=1
frame 2;xlower=0.9;xupper=1.3;ylower=0.5;yupper=0.9

dgraph[keywindow=2;keydescription='Objects and Variables';\
endaction=continue;title='Prediction PCA Biplot']\
pyaxes[1...noofvars],units[2];\
pxaxes[1...noofvars],units[1];description=varsname

variate[nvalues=2]ypoints
variate[nvalues=2]xpoints
scalar ypoint
scalar upoint[1,2]

dread[p=*;cursortype=4]y=ypoints;x=xpoints;ygiven=units[2];xgiven=units[1];\
ysave=unitsave[2];xsave=unitsave[1]
equate[oldf=!(-1,1)]ypoints;ypoint
equate unitsave[1];upoint[1]
equate unitsave[2];upoint[2]

for ind=1...nvars
model marker[ind]
fit [p=*]pyaxes[ind]
rkeep estimates=ps[ind]
endfor

scalar gradient[1...nvars]
scalar cons[1...nvars]
equate[oldf=!(-1,1)]ps[];gradient[]
```

```
equate ps[];cons[]

variate[nvalues=noofvars;values=1...noofvars]ident1
calc ident1=ident1*-1
variate[nvalues=noofvars;values=#noofvars(0.1)]ident2

text[nvalues=noofvars]names
equate varsname;names

axes[equal=no]3;style=none
frame 3;xlower=0.9;xupper=1.3;ylower=0;yupper=0.5
pen 1;m=point;symbol=2;size=1;rotation=-60;labels=names

dgraph[keywindow=0;window=3;title='Variable Selection';\
endaction=continue;screen=keep]ident1;ident2

variate[nvalues=1]yyy
variate[nvalues=1]xxx

dread[print=*;window=3]y=yyy;x=xxx;ygiven=ident1;xgiven=ident2

calculate yyy=round(yyy)
calc yyy=yyy*-1

scalar varsel
equate yyy;varsel

calc predval=(gradient[varsel]*ypoint)*stddev[varsel]+mean(ydata[varsel])
calc varno=nvalues(varname)
variate[nvalues=varno;values=1...varno]varcount

restrict varname;condition=varcount.eq.varsel
restrict unitname;\
condition=(upoint[1].eq.units[1]).and.(upoint[2].eq.units[2])

print 'the predicted value for variable'
print[iprint=*]varname
print 'and unit'
print[iprint=*]unitname
print 'is'
print[iprint=*]predval

restrict unitname
restrict varname

print unitname,ydata[varsel],data[varsel]

endif


"marker 7"

calc govars=vrhovar[1]**2+vrhovar[2]**2
text[nvalues=nvars]Variable
equate[oldf=!(#nvars,-1)]varname;Variable

scalar lrvscal[1...nvars]
equate lrv[2];lrvscal
calc gounits=((lrvscal[1]+lrvscal[2])/lrv[3])*100

print[iprint=*]'The first two principal components explain',gounits
print 'percent of the total variation in the data'

print 'The adequacy of fit of the variables in two dimensions is'
print[iprint=*]Variable,govars

variate[nvalues=nvars]ev[1...nvars]
equate[oldf=!((1,nnvarsm1)#nvars,-1)]lrv[1];ev
calc evsquare[1...nvars]=ev[]**2

scalar l[1...nvars]
equate lrv[2];l

calc cusum[1...nvars]=evsquare[]*l[]

factor[nvalues=nvars;levels=nvars;values=1...nvars]PrinComp
factor[nvalues=nvars;levels=nvars;values=1...nvars;labels=Variable]vars

pen 1...20;labels=*;brush=16
axes 1;ytitle='Eigenvalue';xtitle='Principal Components'
table[class=PrinComp,vars]cusumtab
equate cusum;cusumtab
dbarchart[title='CUSUM diagram';\
append=yes;keydescription='Variables']\
cusumtab
```

7 Gruppenanalysemodell und Gamma-q-q-Plots

```
"Gruppenanlysemodell und Gamma-q-q-Plots"


"OPENING DATA SET"

open 'c:\\phd\\kennzahl\\data\\gs_kn5.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b


"DEFINITION OF DATA POINTER"

pointer[values=allgawp,spezp,lohnqp,lohnak,heizqm,\
eqm,glasqm,glasqmak,\
fkp,anvermp,\
beinkp,beinkak,beinkeqm,kapkoef,rdiffp,rentkoef]data


"DEFINITION OF FACTOR POINTER"
pointer[values=qm1_shn,fregion,fjahr]fdata


"USE OF COVARIANCE SSPM OR CORRELATION MATRIX"
"use corr; variance and  ssp have to be checked"

text[nvalues=1;value='corr']usemat


"NORMAL OR ROBUST PCA ESTIMATORS"
"use robust or normal"

text[nvalues=1;value='robust']estim

"_____________________________________________"

"preliminaries"

calc nvars=nvalues(data)
calc nvals=nvalues(data[1])
calc fnvars=nvalues(fdata)
variate[nvalues=nvals;values=1...nvals]case
facproduct fdata;product=comb
calc lcomb=nlevels(comb)

"_____________________________________________"

"forming and storing subsets"

for ind=1...lcomb
subset[condition=comb.eq.ind]\
data[1...nvars],case;subset[1...nvars][ind],subcase[ind]
pointer[values=subset[][ind]]subspace[ind]
endfor

"_____________________________________________"


"performing pca on normal estimates"

if estim .eqs. 'normal'

for ind=1...lcomb
pcp[p=roots;m=#usemat]subspace[ind];lrv=l[ind]
lrvscree l[ind]
endfor


"performing pca on robust estimates"

elsif estim .eqs. 'robust'

for ind =1...lcomb
robsspm[p=outlier]subspace[ind];sspm=subsspm[ind]
calc subvars=nvalues(subspace[ind][1])
variate[nvalues=subvars;values=1...subvars]case_no
print case_no,subcase[ind]
pcp[p=roots;m=#usemat]subsspm[ind];lrv=l[ind]
lrvscree l[ind]
endfor

endif
```

```
"storing data"

open 'c:\\phd\\kennzahl\\data\\gs_subs1.glg';c=5;f=b
store[c=5;l=i;m=replace]l,subspace,subcase,\
nvars,nvals,fnvars,comb,lcomb,estim,usemat
close 5;f=b


"deletion"

delete[redefine=yes;l=e]subspace,subcase,subsspm,\
nvars,nvals,fnvars,comb,lcomb,estim,usemat

"_____________________________________________"


"groupwise analysis"


for ind =1...lcomb

print '*****      Analysis of Subgroup',ind, '*****' ;decimals=0


"preliminaries"

calc subvals[ind]=nvalues(subspace[ind][1])
calc subvars[ind]=nvalues(subspace[ind])
variate[nvalues=subvals[ind]]case_no[ind]
equate subcase[ind];case_no[ind]


"NAMING VARIABLES"

calc allgawp[ind]=subset[1][ind]
calc spezp[ind]=subset[2][ind]
calc lohnqp[ind]=subset[3][ind]
calc lohnak[ind]=subset[4][ind]
calc heizqm[ind]=subset[5][ind]
calc eqm[ind]=subset[6][ind]
calc glasqm[ind]=subset[7][ind]
calc glasqmak[ind]=subset[8][ind]
calc fkp[ind]=subset[9][ind]
calc anvermp[ind]=subset[10][ind]
calc beinkp[ind]=subset[11][ind]
calc beinkak[ind]=subset[12][ind]
calc beinkeqm[ind]=subset[13][ind]
calc kapkoef[ind]=subset[14][ind]
calc rdiffp[ind]=subset[15][ind]
calc rentkoef[ind]=subset[16][ind]


"REDEFINITION OF DATA POINTER"

pointer[values=allgawp[ind],spezp[ind],lohnqp[ind],lohnak[ind],heizqm[ind],\
eqm[ind],glasqm[ind],glasqmak[ind],\
fkp[ind],anvermp[ind],\
beinkp[ind],beinkak[ind],beinkeqm[ind],kapkoef[ind],rdiffp[ind],rentkoef[ind]]data[ind]


"VARIABLE NAMES AND UNITS DESCRIPTION"
text[values='allgawp','spezp','lohnqp','lohnak','heizqm',\
'eqm','glasqm','glasqmak',\
'fkp','anvermp',\
'beinkp','beinkak','beinkeqm','kapkoef','rdiffp','rentkoef']varname


"NUMBER OF DIMENSIONS TO LOOK AT"
scalar[value=4]noofdims


"UNIT NAMES"
ftext case_no[ind];unitname[ind]

"_____________________________________________"

"Group comparisons on robust estimates"


"standardisation"

if usemat .eqs. 'corr'
text[nvalues=1;values='yes']standard
```

```
else
text[nvalues=1;values='no']standard
endif

if standard .eqs. 'yes'

calculate variance[ind][1...nvars]=var(data[ind][])
calculate stddev[ind][1...nvars]=sqrt(variance[ind][])

for j=1...subvars[ind]

if stddev[ind][j]>0
calculate zdata[ind][j]=(data[ind][j]-mean(data[ind][j]))/stddev[ind][j]
else
calculate zdata[ind][j]=('missing')
endif

endfor

calc data[ind][]=zdata[ind][]

elsif standard .eqs. 'no'
calc data[ind][]=data[ind][]

endif


"robust pca"

robsspm[p=outliers]data[ind];sspm=subsspm[ind]
variate[nvalues=subvals[ind];values=1...subvals[ind]]case[ind]
print case[ind],unitname[ind]
pcp[m=#usemat;nroots=noofdims]subsspm[ind];lrv=l[ind]
print l[ind][]
endfor


"deletion"

delete[redefine=yes;l=e]l,lcomb,nvars,nvals,subvars,subvals,\
noofdims,varname,fdata,flevels,estim,usemat


"calculating average vectors"

calc nvarssq=nvars*nvars
matrix[r=nvars;c=nvars;values=#nvarssq(0)]H

for ind=1...lcomb

calc LtLt[ind]=rtproduct(l[ind][1];l[ind][1])
calc H=H+LtLt[ind]

endfor

variate[nvalues=nvars]varH[1...nvars]
equate H;varH
pcp[nroots=noofdims]varH;lrv=hl
svd H;singular=s
variate[nvalues=noofdims]sscos
equate s;sscos

for k=1...noofdims

calc b[k]=hl[1]$[*;k]

endfor


"calculating angles"

for k=1...noofdims
for ind=1...lcomb

calc deltat[k][ind]=arccos(sqrt(t(b[k])*+LtLt[ind]*+b[k]))*57.29578

endfor
endfor


"preparing the printout"

for k=1...noofdims
variate[nvalues=lcomb;values=deltat[k][]]delta[k]
endfor
```

```
variate[nvalues=lcomb;values=1...lcomb]grp
ftext grp;group

print '***   Between-Groups Comparison of Principal Components   ***'
print 'average component loadings b that minimize V'
print 'directions closest to each subspace'
print varname,b[]

print 'angles formed by each group with each direction'
print group,delta[]

print 'sum of squared cosines'
print[orientation=across]sscos

print 'the groups are defined by'
print[orientation=across;iprint=*]fdata
print 'with levels'
print[iprint=*]flevels[]

"deletion"

delete[redefine=yes]

"_____"

"gamma-q-q-plots"

"DEFINITION OF EIGENVECTOR TO BE COMPARED"

scalar[value=1]ev


"forming variates from roots for boxplots"

variate[nvalues=lcomb]evals[1...nvars]
variate[nvalues=nvars]lvar[1...lcomb]
matrix[r=lcomb;c=nvars]lmat

for ind=1...lcomb
equate l[ind][2];lvar[ind]
endfor

equate lvar;lmat
calc tlmat=t(lmat)
equate tlmat;evals


"drawing the boxplots"

variate[nvalues=nvars;values=(1...nvars)]xlab
variate[nvalues=1]xsca[1...nvars]
equate xlab;xsca

ftext xsca[];pcs[1...nvars]

boxplot[title='Boxplots of eigenvalues of subgroups for \
all principal components';\
axistitle='Eigenvalue']evals[];boxlabels=pcs[]


"finding the typical vector"

calc nsets=nvalues(l)
calc nvars=sqrt(nvalues(l[1][1]))

matrix[r=nvars;c=1]evperset[1...nsets]

for ind=1...nsets
calc evperset[ind]=l[ind][1]$[*;ev]
endfor

calc tevpers[1...nsets]=t(evperset[1...nsets])
calc evpers2[1...nsets]=evperset[]*+tevpers[]

calc E=evpers2[1]

for ind=2...nsets
calc E=E+evpers2[ind]
endfor

pcp E;lrv=lall

matrix[r=nvars;c=1]evtyp
calc evtyp=lall[1]$[*;1]
```

```
"calculating the dissimilarities"

for ind=1...nsets
calc dissimm[ind]=t(evtyp-evperset[ind])*+(evtyp-evperset[ind])
calc dissimp[ind]=t(evtyp+evperset[ind])*+(evtyp+evperset[ind])
variate[nvalues=2]dissimv[ind]
equate !P(dissimm[ind],dissimp[ind]);dissimv[ind]
calc dissim[ind]=min(dissimv[ind])
endfor

variate[nvalues=nsets]dist
equate dissim;dist


"fitting the gamma distribution to the distances"

distribution[dist=gamma]dist;parameters=param
scalar eta,lambda
equate param;!P(eta,lambda)

variate[nvalues=nsets;values=1...nsets]index
calc p=(index-0.375)/(nsets+0.25)
calc m=eta
calc d=1/lambda
calc gamma=edgamma(p;m;d)

sort dist,index
sort gamma


"drawing the qq gamma plot"

ftext index;groups

variate[nvalues=1;values=0]null[1,2]
variate[nvalues=1]gamax[1,2]
calc gamax[1,2]=max(gamma)
variate[nvalues=2]linepoin[1,2]
append[newvector=linepoin[1]]null[1],gamax[1]
append[newvector=linepoin[2]]null[2],gamax[2]

pen 1;labels=groups;size=0.75;symbol=2
pen 2;m=line;linestyle=1;symbol=0


axes[equal=scale] 1;xtitle='Gamma quantiles';ytitle='Ordered squared distances';\
style=box
dgraph [title='Gamma q-q plot, ';key=0]dist,linepoin[2];gamma,linepoin[1]
```

8 Stabilitätsprüfung

```
"Stabilitätzsprüfung Merkmale"


"READING DATA"

UNIT [*]
FILEREAD [PRINT=summary,groups,comments,firstline;\
name='C:/personal/CYCLAMEN/DATA/ED_CY1_3.DAT';\
MISSING='*'; SEPARATOR=' '; IMETHOD=read] FGROUPS=no

pointer[values= sub_ee  ,    abs_n1 ,   groe_g10 ,    men_g50 ,    sf1_mod,\
sf2_mod  ,  bew1_kop  ,  bew2_kop ,    kre_wes ,    sub_and  ,    abs_g1,\
groe_w10,     men_u50  ,   sf1_alt ,    sf2_alt,    bew1_fus,    bew2_fus,\
kre_ost]data

ftext betrieb;Betrieb


"preliminaries"

calc nvals=nvalues(data[1])
calc nvars=nvalues(data)
calc nvarsm1=nvars-1
calc nvalsm1=nvals-1
calc nvarsp1=nvars+1
calc n=nvals*nvars
calc nmnvars=n-nvars


"DESCRIPTION OF VARIABLES"

text[nvalues=nvars;value=ee  ,vm1 , fg10,    mg50 ,    sf1_m,\
     sf2_m  ,  bw1_k  ,  bw2_k ,    wes ,    subs  ,    vmg1,\
    fw10,     mw50  ,   sf1_a,    sf2_a,    bw1_f,    bw2_f,\
     ost]varnames


"correspondence analysis of full data matrix"

matrix[r=nvals;c=nvars]datamat
matrix[r=nvars;c=nvals]tmat
equate data;tmat
calc datamat=t(tmat)
corresp datamat;rowscore=rows;colscore=cols;\
roots=eigen_f;rowinertia=rowin_f;colinertia=colin_f
variate[nvalues=nvals]rowf[1,2]
variate[nvalues=nvars]colf[1,2]
calc rowf[1,2]=rows$[*;1,2]
calc colf[1,2]=cols$[*;1,2]


"row reduction"

"Generate sorting factors"

factor[nvalues=n;levels=nvars]f1
factor[nvalues=n;levels=nvals]f2
generate f1,f2


"Create row reduced data set"

variate[nvalues=nmnvars]redvar[1...nvals]
for i=1...nvals
variate[nvalues=n]datavar
equate data;datavar
subset[condition=f2.ne.i]datavar
calc redvar[i]=datavar
equate data;datavar
endfor


"Create row reduced data matrices"

matrix[r=nvars;c=nvalsm1]trredmat[1...nvals]
matrix[r=nvalsm1;c=nvars]rredmat[1...nvals]
equate redvar;trredmat
calc rredmat[]=t(trredmat[])


"correspondence analysis of reduced data matrix"
```

```
for ind=1...nvals
corresp rredmat[ind];rowscore=row[ind];\
colscore=col[ind];roots=eigen_r[ind];\
rowinertia=rowin_r[ind];colinertia=colin_r[ind]
calc rowvar[ind][1,2]=row[ind]$[*;1,2]
calc colvar[ind][1,2]=col[ind]$[*;1,2]
endfor


"first deletion"

delete[redefine=yes;l=exclusive]varnames,betrieb,Betrieb,nvals,\
nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,cols,col,rows,row,eigen_f,eigen_r,colin_f,colin_r,\
rowin_f,rowin_r,rowf,colf,rowvar,colvar


"results from simple ca"

calc nnvalsm1=(nvals-1)*-1
calc dp_nvars=nvars


"calcs for output"

calc rootsumt=sum(eigen_f)
scalar ro12[1...2]
equate eigen_f;ro12
calc varex1=(ro12[1]/rootsumt)*100
calc varex2=(ro12[2]/rootsumt)*100
calc nvalro=nvalues(eigen_f)

scalar rox[1...nvalro]
equate eigen_f;rox


"absolute contributions"

calc colinf12[1...2]=colin_f$[*;1,2]
calc colabcon[1]=colinf12[1]/ro12[1]
calc colabcon[2]=colinf12[2]/ro12[2]

calc rowinf12[1...2]=rowin_f$[*;1,2]
calc rowabcon[1]=rowinf12[1]/ro12[1]
calc rowabcon[2]=rowinf12[2]/ro12[2]

calc rowcon1[1...nvalro]=(rowin_f$[*;1...nvalro])/rox[1...nvalro]
calc colcon1[1...nvalro]=(colin_f$[*;1...nvalro])/rox[1...nvalro]

calc rowcon2[1...nvalro]=(rowin_f$[*;1...nvalro])
calc colcon2[1...nvalro]=(colin_f$[*;1...nvalro])

calc Unit_CON[1,2]=rowabcon[]*100

calc Var_CON[1,2]=colabcon[]*100


"relative contributions"

variate[nvalues=nvals]rowinvar[1...nvalro]
equate rowcon2;rowinvar
variate[nvalues=dp_nvars]colinvar[1...nvalro]
equate colcon2;colinvar

calc inj=vsum(colinvar)
calc ink=vsum(rowinvar)

calc Var_COR[1...nvalro]=colinvar[]/inj
calc Unit_COR[1...nvalro]=rowinvar[]/ink

calc Var_QLT=Var_COR[1]+Var_COR[2]
calc Unit_QLT=Unit_COR[1]+Unit_COR[2]


"printed output"

print 'roots of full data matrix'
print eigen_f
print[iprint=*]'percentage of inertia explained \
through first principal axis',varex1
print[iprint=*]'percentage of inertia explained \
through second principal axis',varex2

print  'absolue unit contribution to the first two dimensions in %'
```

```
print Betrieb,Unit_CON[1],Unit_CON[2]

print  'absolue variable contribution to the first two dimensions in %'
print varnames,Var_CON[1],Var_CON[2]

print 'relative contributions and quality \
of the two-dimensional display of the units'
print Betrieb,Unit_QLT,Unit_COR[1],Unit_COR[2]

print 'relative contributions and quality \
of the two-dimensional display of the variables'
print varnames,Var_QLT,Var_COR[1],Var_COR[2]


"second deletion"

delete[redefine=yes;l=exclusive]varnames,betrieb,Betrieb,\
nvals,nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,cols,col,rows,row,eigen_f,eigen_r,colin_f,colin_r,rowin_r


"results from simple ca for reduced data matrix"

calc nnvalsm1=(nvals-1)*-1
calc dp_nvars=nvars


"calcs for output"

for ind=1...nvals
calc rrootsum[ind]=sum(eigen_r[ind])
scalar rro12[ind][1...2]
equate eigen_r[ind];rro12[ind]
calc rvarex1[ind]=(rro12[ind][1]/rrootsum[ind])*100
calc rvarex2[ind]=(rro12[ind][2]/rrootsum[ind])*100
calc rnvalro[ind]=nvalues(eigen_r[ind])

scalar rrox[ind][1...rnvalro[ind]]
equate eigen_r[ind];rrox[ind]


"absolute contributions"

calc rcolin12[ind][1...2]=colin_r[ind]$[*;1,2]
calc rcolabco[ind][1]=rcolin12[ind][1]/rro12[ind][1]
calc rcolabco[ind][2]=rcolin12[ind][2]/rro12[ind][2]

calc rrowin12[ind][1...2]=rowin_r[ind]$[*;1,2]
calc rrowabco[ind][1]=rrowin12[ind][1]/rro12[ind][1]
calc rrowabco[ind][2]=rrowin12[ind][2]/rro12[ind][2]

calc rrowcon1[ind][1...rnvalro[ind]]=(rowin_r[ind]\
$[*;1...rnvalro[ind]])/rrox[ind][1...rnvalro[ind]]
calc rcolcon1[ind][1...rnvalro[ind]]=(colin_r[ind]\
$[*;1...rnvalro[ind]])/rrox[ind][1...rnvalro[ind]]

calc rrowcon2[ind][1...rnvalro[ind]]=(rowin_r[ind]$[*;1...rnvalro[ind]])
calc rcolcon2[ind][1...rnvalro[ind]]=(colin_r[ind]$[*;1...rnvalro[ind]])

calc U_CON[ind][1,2]=rrowabco[ind][]*100

calc V_CON[ind][1,2]=rcolabco[ind][]*100

variate[nvalues=nvalsm1]rbetrieb[ind]
subset[condition=betrieb.ne.ind]betrieb;rbetrieb[ind]


"printed output 1"

print 'roots of reduced data matrix'
print eigen_r[ind]
print[iprint=*]'percentage of inertia explained \
through first principal axis',rvarex1[ind]
print[iprint=*]'percentage of inertia explained \
through second principal axis',rvarex2[ind]

print  'absolue unit contribution to the first two dimensions in %'
print rbetrieb[ind],U_CON[ind][1],U_CON[ind][2]

print  'absolue variable contribution to the first two dimensions in %'
print varnames,V_CON[ind][1],V_CON[ind][2]


"third deletion"
```

```
delete[redefine=yes;l=exclusive]varnames,betrieb,Betrieb,\
dp_nvars,nnvalsm1,nvals,nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,cols,col,rows,row,eigen_r,colin_r,rowin_r,rrowcon2,rcolinva,\
rcolcon2,rrowinva,rnvalro,rbetrieb,rvarex1,rvarex2

endfor


"forth deletion"

delete[redefine=yes;l=exclusive]varnames,betrieb,Betrieb,\
dp_nvars,nnvalsm1,nvals,nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,cols,col,rows,row,rrowcon2,rcolcon2,rnvalro,rbetrieb,rvarex1,rvarex2


"relative contributions"

for ind=1...nvals
variate[nvalues=nvalsm1]rrowinva[ind][1...rnvalro[ind]]
equate rrowcon2[ind];rrowinva[ind]
variate[nvalues=dp_nvars]rcolinva[ind][1...rnvalro[ind]]
equate rcolcon2[ind];rcolinva[ind]

calc rinj[ind]=vsum(rcolinva[ind])
calc rink[ind]=vsum(rrowinva[ind])

calc V_COR[ind][1...rnvalro[ind]]=rcolinva[ind][]/rinj[ind]
calc U_COR[ind][1...rnvalro[ind]]=rrowinva[ind][]/rink[ind]

calc V_QLT[ind]=V_COR[ind][1]+V_COR[ind][2]
calc U_QLT[ind]=U_COR[ind][1]+U_COR[ind][2]


"printed output 2"

print 'relative contributions and quality of \
the two-dimensional display of the units'
print rbetrieb[ind],U_QLT[ind],U_COR[ind][1],U_COR[ind][2]

print 'relative contributions and quality of \
the two-dimensional display of the variables'
print varnames,V_QLT[ind],V_COR[ind][1],V_COR[ind][2]


"fifth deletion"

delete[redefine=yes;l=exclusive]varnames,betrieb,Betrieb,\
dp_nvars,nnvalsm1,nvals,nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,cols,col,rows,row,eigen_r,colin_r,rowin_r,rrowcon2,rcolinva,\
rcolcon2,rrowinva,rnvalro,rbetrieb,rvarex1,rvarex2

endfor


"sixth deletion"

delete[redefine=yes;l=exclusive]varnames,nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,col,cols,row,rows


"procrustes rotation"

for ind=1...nvals
rotate cols;col[ind];xout=fixout[ind];yout=fitout[ind];\
residuals=resi[ind]
variate[nvalues=nvars]con_wo[ind][1,2]
calc con_wo[ind][1,2]=fitout[ind]$[*;1,2]
calc confix[ind][1,2]=fitout[ind]$[*;1,2]
endfor


"seventh deletion"

delete[redefine=yes;l=exclusive]varnames,nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,col,row,con_wo,confix,resi


"rotated jacknifed configurations"

variate[nvalues=nvars;values=1...nvars]variable
ftext variable;Variable
axes[equal=scale] 1;style=box;xtitle='first principal axis';\
ytitle='second principal axis'
```

```
pen 1;labels=varnames;symbol=2;size=0.75

for ind=1...nvals
dgraph[title='Configuration of variables without unit n']\
con_wo[ind][2];con_wo[ind][1]
endfor


"dotplot of residuals"

variate[nvalues=nvars]residual[1...nvals]
equate resi;residual

for ind=1...nvals
calc  resimax1[ind]=max(residual[ind])
variate[nvalues=nvals]resimax2
equate resimax1;resimax2
calc resimax=max(resimax2)
endfor

for ind=1...nvals

pen 1;labels=*
if resimax1[ind]<0.05
axes[equal=no]1;xtitle=*;\
ytitle='variables';xlower=0;xupper=0.05
else
axes[equal=no]1;xtitle=*;\
ytitle='variables';xlower=0;xupper=1
endif
dotplot[g=h;title='Residuals\
 of rotated jackknifed configuration of variables']\
varnames;residual[ind]

endfor


"eigth deletion"

delete[redefine=yes;l=exclusive]varnames,nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,con_wo,confix


"preparing convex hulls"

for ind =2...nvals
append con_wo[1][1],con_wo[ind][1]
append con_wo[1][2],con_wo[ind][2]
endfor

variate[nvalues=n;values=(1...nvars)#nvals]sortlist

for ind=1...nvars
variate[nvalues=nvals]peelvar[ind][1,2]
subset[condition=sortlist.eq.ind]\
con_wo[1][1],con_wo[1][2];peelvar[ind][1,2]
endfor


"ninth deletion"

delete[redefine=yes;l=exclusive]varnames,nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,confix,peelvar

for ind=1...nvars
convexhull peelvar[ind][2];peelvar[ind][1];\
yhull=yhull[ind];xhull=xhull[ind]
endfor


"drawing convex hulls"

pen 1;symbol=2;labels=varnames;size=0.5
axes[equal=scale]1;xtitle='first principal axis';\
ytitle='second principal axis';style=box;xlower=*;xupper=*
pen 2...nvarsp1;symbol=0;method=line;\
linestyle=1;join=given;colour=4
dgraph[key=0;title='Convex hulls of jackknifed \
configurations of variables']confix[1][2],yhull[1...nvars];\
confix[1][1],xhull[1...nvars]


"Stabilitätzsprüfung Objekte"
```

```
"READING DATA"

UNIT [*]
FILEREAD [PRINT=summary,groups,comments,firstline;\
name='C:/personal/CYCLAMEN/DATA/ED_CY1_3.DAT';\
MISSING='*'; SEPARATOR=' '; IMETHOD=read] FGROUPS=no

pointer[values= sub_ee  ,    abs_n1 ,   groe_g10 ,    men_g50 ,    sf1_mod,\
sf2_mod  ,  bew1_kop  ,  bew2_kop ,    kre_wes ,    sub_and  ,    abs_g1,\
groe_w10,     men_u50  ,   sf1_alt ,    sf2_alt,    bew1_fus,    bew2_fus,\
kre_ost]data

ftext betrieb;Betrieb


"preliminaries"

calc nvals=nvalues(data[1])
calc nvars=nvalues(data)
calc nvarsm1=nvars-1
calc nvalsm1=nvals-1
calc nvarsp1=nvars+1
calc n=nvals*nvars
calc nmnvars=n-nvars

for ind =1...nvals
variate[nvalues=nvalsm1]rbetrieb[ind]
subset[condition=betrieb.ne.ind]betrieb;rbetrieb[ind]
ftext rbetrieb[ind];rBetrieb[ind]
endfor


"DESCRIPTION OF VARIABLES"

text[nvalues=nvars;value=ee  ,vm1 , fg10,    mg50 ,    sf1_m,\
     sf2_m  ,  bw1_k  ,  bw2_k ,    wes ,    subs  ,    vmg1,\
    fw10,     mw50  ,   sf1_a,    sf2_a,    bw1_f,    bw2_f,\
     ost]varnames


"correspondence analysis of full data matrix"

matrix[r=nvals;c=nvars]datamat
matrix[r=nvars;c=nvals]tmat
equate data;tmat
calc datamat=t(tmat)
corresp datamat;rowscore=rows;colscore=cols;\
roots=eigen_f;rowinertia=rowin_f;colinertia=colin_f
variate[nvalues=nvals]rowf[1,2]
variate[nvalues=nvars]colf[1,2]
calc rowf[1,2]=rows$[*;1,2]
calc colf[1,2]=cols$[*;1,2]


"row reduction"

"Generate sorting factors"

factor[nvalues=n;levels=nvars]f1
factor[nvalues=n;levels=nvals]f2
generate f1,f2


"Create row reduced data set"

variate[nvalues=nmnvars]redvar[1...nvals]
for i=1...nvals
variate[nvalues=n]datavar
equate data;datavar
subset[condition=f2.ne.i]datavar
calc redvar[i]=datavar
equate data;datavar
endfor


"Create row reduced data matrices"

matrix[r=nvars;c=nvalsm1]trredmat[1...nvals]
matrix[r=nvalsm1;c=nvars]rredmat[1...nvals]
equate redvar;trredmat
calc rredmat[]=t(trredmat[])
```

```
"correspondence analysis of reduced data matrix"

for ind=1...nvals
corresp rredmat[ind];rowscore=row[ind];\
colscore=col[ind];roots=eigen_r[ind];\
rowinertia=rowin_r[ind];colinertia=colin_r[ind]
calc rowvar[ind][1,2]=row[ind]$[*;1,2]
calc colvar[ind][1,2]=col[ind]$[*;1,2]
endfor


"first deletion"

delete[redefine=yes;l=exclusive]rbetrieb,betrieb,Betrieb,rBetrieb,nvals,\
nvars,nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,rows,row


"sixth deletion"

delete[redefine=yes;l=exclusive]rbetrieb,betrieb,Betrieb,rBetrieb,nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,row,rows


"generating fix row matrices"

variate[nvalues=nvals]rowred[1...nvarsm1]
calc rowred[1...nvarsm1]=rows$[*;1...nvarsm1]
for ind=1...nvals
variate[nvalues=nvalsm1]rrowred[ind][1...nvarsm1]
subset[condition=betrieb.ne.ind]rowred[1...nvarsm1];\
rrowred[ind][1...nvarsm1]
endfor

calc nnvalsm2=-1*(nvals-2)
matrix[r=nvalsm1;c=nvarsm1]rrows[1...nvals]

for ind =1...nvals
equate[oldf=!((1,nnvalsm2)#nvarsm1,-1)] rrowred[ind];rrows[ind]
endfor


"seventh deletion"

delete[redefine=yes;l=exclusive]rbetrieb,Betrieb,rBetrieb,betrieb,\
nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,row,rows,rrows


"procrustes rotation"

for ind=1...nvals
rotate rrows[ind];row[ind];xout=fixout[ind];yout=fitout[ind];\
residuals=resi[ind]
variate[nvalues=nvalsm1]r_con_wo[ind][1,2]
variate[nvalues=nvalsm1]r_confix[ind][1,2]
calc r_con_wo[ind][1,2]=fitout[ind]$[*;1,2]
calc r_confix[ind][1,2]=fitout[ind]$[*;1,2]
endfor


"rotated jacknifed configurations"

axes[equal=scale] 1;style=box;xtitle='first principal axis';\
ytitle='second principal axis'

for ind=1...nvals
pen 1;labels=rBetrieb[ind];symbol=1;size=0.75
dgraph[title='Configuration of units without unit n']\
r_con_wo[ind][2];r_con_wo[ind][1]
endfor


"dotplot of residuals"

variate[nvalues=nvalsm1]residual[1...nvals]
equate resi;residual

for ind=1...nvals
calc  resimax1[ind]=max(residual[ind])
variate[nvalues=nvals]resimax2
equate resimax1;resimax2
calc resimax=max(resimax2)
```

```
endfor

for ind=1...nvals

pen 1;labels=*
if resimax1[ind]<0.01
axes[equal=no]1;xtitle=*;\
ytitle='units';xlower=0;xupper=0.01
else
axes[equal=no]1;xtitle=*;\
ytitle='units';xlower=0;xupper=0.2
endif
dotplot[g=h;title='Residuals\
 of rotated jackknifed configuration of units']\
rBetrieb[ind];residual[ind]

endfor


"eigth deletion"

delete[redefine=yes;l=exclusive]Betrieb,rbetrieb,betrieb,rBetrieb,\
nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,r_con_wo,r_confix


"preparing convex hulls"

for ind =2...nvals
append r_con_wo[1][1],r_con_wo[ind][1]
append r_con_wo[1][2],r_con_wo[ind][2]
append r_confix[1][1],r_confix[ind][1]
append r_confix[1][2],r_confix[ind][2]
endfor

calc rn=nvals*nvalsm1
variate[nvalues=rn]sortlist
equate rbetrieb;sortlist
factor[nvalues=rn;levels=!(1...nvals)] f_sort
equate sortlist;f_sort

tabulate[class=f_sort] r_confix[1][1];means=rfix[1][1]
tabulate[class=f_sort] r_confix[1][2];means=rfix[1][2]

variate[nvalues=nvals]rfixvar[1][1,2]
equate rfix[1][1];rfixvar[1][1]
equate rfix[1][2];rfixvar[1][2]

for ind=1...nvals
variate[nvalues=nvalsm1]peelvar[ind][1,2]
subset[condition=sortlist.eq.ind]\
r_con_wo[1][1],r_con_wo[1][2];peelvar[ind][1,2]
endfor


"ninth deletion"

delete[redefine=yes;l=exclusive]rbetrieb,Betrieb,betrieb,rBetrieb,\
nvals,nvars,\
nvarsm1,nvalsm1,nvarsp1,n,nmvars,\
data,peelvar,rfixvar


"calculating convex hulls"

for ind=1...nvals
convexhull peelvar[ind][2];peelvar[ind][1];\
yhull=yhull[ind];xhull=xhull[ind]
endfor


"drawing convex hulls"

calc nvalsp1=nvals+1
pen 1;symbol=1;labels=Betrieb;size=0.5
axes[equal=scale]1;xtitle='first principal axis';\
ytitle='second principal axis';style=box;xlower=*;xupper=*
pen 2...nvalsp1;symbol=0;method=line;\
linestyle=1;join=given;colour=4
dgraph[key=0;title='Convex hulls of jackknifed \
configurations of units']rfixvar[1][2],yhull[1...nvals];\
rfixvar[1][1],xhull[1...nvals]
```

9 Genstat-Menus zur Ergänzung der formalen Begriffsanalyse

```
"Genstat-Menus zum Start aus Access in Analyse hierarchischer Liniendiagramme"


"question how to start Genstat"

QUESTION [PREAMBLE=!t('How do you want to start Genstat');\
  RESPONSE=_start; \
  MODE=t; DEFAULT='s'] VALUES='s','db1','db2'; CHOICE= \
  'standard', \
  'database kennzahlen',\
  'database cyclamen'


"Nelders proc for summary statistics"

proc[par=p] 'summ'      " prints summary statistics of variates "
param 'X'

getatt[nval] X; !p(nv)
scal (len,mn,v,md,mnm,mxm)[1...nv]
calc len[]=nval(#X) & mn[]=mean(#X)  & v[]=var(#X)
   &  md[]=med(#X) & mnm[]=min(#X) & mxm[]=max(#X)

print[sq=y;ipr=*] 'length','mean',' var.','median','min.','max.'
for j=1...nv
print[sq=y;ipr=*] (len,mn,v,md,mnm,mxm)[j]
endf

endp


if _start.eq.2

delete[redefine=yes;l=e]

"opening the database"

set[in=*]
open 'c:\\phd\\kennzahl\\fba\\data\\kennzahl.glg';c=5;f=b
retrieve[c=5;l=a]
close 5;f=b

"reading data from the clipboard"

UNIT [*]
FILEREAD [PRINT=summary,groups,comments,firstline;\
name='C:/phd/kennzahl/fba/schema/cases.txt';\
 MISSING='*'; SEPARATOR=' '; IMETHOD=supply]C1; FGROUPS=no


"graphical environment"

frame 1;xlower=0;xupper=1;ylower=0;yupper=1


"using question to select additional statistics"

QUESTION [PREAMBLE=!t('Further Analysis of Groups',*, \
  'Do you want to see a SCATTERPLOT MATRIX'); RESPONSE=_sourc1; \
  MODE=t; DEFAULT='y'] VALUES='y','n'; CHOICE= \
  'yes', \
  'no'

if _sourc1.eq.1

QUESTION [PREAMBLE=!t('Further Analysis of Groups - SCATTERPLOT MATRIX',*, \
  'Select the variables to look at'); RESPONSE=_scvar; \
  list=yes;DECLARED=yes; TYPE=variate; PRESENT=yes]

endif


QUESTION [PREAMBLE=!t('Further Analysis of Groups',*, \
  'Do you want to see a BOXPLOT'); RESPONSE=_sourc3; \
  MODE=t; DEFAULT='y'] VALUES='y','n'; CHOICE= \
  'yes', \
  'no'

if _sourc3.eq.1
```

```
QUESTION [PREAMBLE=!t('Further Analysis of Groups - BOXPLOTS',*, \
  'Select the variables to look at'); RESPONSE=_bvar; \
  list=yes;DECLARED=yes; TYPE=variate; PRESENT=yes]

endif


QUESTION [PREAMBLE=!t('Further Analysis of Groups',*, \
  'Do you want to look at some SUMMARY STATISTICS'); RESPONSE=_sourc4; \
  MODE=t; DEFAULT='y'] VALUES='y','n'; CHOICE= \
  'yes', \
  'no'

if _sourc4.eq.1


QUESTION [PREAMBLE=!t('Further Analysis of Groups - SUMMARY STATISTICS',*, \
  'Select the variables to look at'); RESPONSE=_svar; \
  list=yes;DECLARED=yes; TYPE=variate; PRESENT=yes]

endif


QUESTION [PREAMBLE=!t('Further Analysis of Groups',*, \
  'Do you want to see a DOTPLOT'); RESPONSE=_sourc2; \
  MODE=t; DEFAULT='y'] VALUES='y','n'; CHOICE= \
  'yes', \
  'no'


if _sourc2.eq.1

QUESTION [PREAMBLE=!t('Further Analysis of Groups - DOTPLOTS',*, \
  'Select the variables to look at'); RESPONSE=_dvar; \
  list=yes;DECLARED=yes; TYPE=variate; PRESENT=yes]

endif


"subsets for making the scatterplot matrix"

if _sourc1.eq.1
restrict _scvar[];condition=case.in.C1

"making the scatterplot-matrix"

dscatter _scvar[]
restrict _scvar[]

endif


"subsets for making boxplots"

if _sourc3.eq.1
restrict _bvar[];condition=case.in.C1

"making the boxplot"

calc nvars=nvalues(_bvar)
for ind=1...nvars
boxplot [g=h;m=schematic;title='boxplot for the selected group']_bvar[ind]
endfor
restrict _bvar[]

endif


"summary statistics for the whole data set"

if _sourc4.eq.1

print 'summary statistics for the whole data set'
print _svar
summ _svar[]


"summary statistics for the selected group"

restrict _svar[];condition=case.in.C1
print 'summary statistics for the selected group'
print _svar
summ _svar[]
restrict _svar[]
```

```
endif


"subsets for making dotplots"

if _sourc2.eq.1

ftext case;betrieb
calc nvars=nvalues(_dvar)
for ind=1...nvars
calc data[ind]=_dvar[ind]
subset[condition=case.in.C1]data[ind];_dvar[ind]
endfor
subset[condition=case.in.C1]betrieb;betriebe


"making the dotplot"

for ind =1...nvars
calc mindvar[ind]=min(_dvar[ind])
axes 1;xlower=mindvar[ind]
dotplot [g=h;direction=ascending;title='dotplot  for the selected group']\
betriebe;_dvar[ind]
endfor
axes 1;xlower=*

endif


"----------------------------------------------------------------------"
elsif _start.eq.3

delete[redefine=yes;l=e]

"opening the database"

print 'cyclamen database start-up file not yet defined'


"----------------------------------------------------------------"
elsif _start.eq.1

delete[redefine=yes;l=e]

"opening the database"

print 'standard start-up'

endif
```